Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Desenvolvimento de um modelo para construção de mapas genéticos em autopoliploides, com aplicações em cana-de-açúcar

Marcelo Mollinari

Tese apresentada para obtenção do título de Doutor em Ciências. Área de concentração: Genética e Melhoramento de Plantas

Piracicaba
2012

Marcelo Mollinari
Engenheiro Agrônomo

Desenvolvimento de um modelo para construção de mapas genéticos em autopoliploides, com aplicações em cana-de-açúcar

Orientador:
Prof. Dr. **ANTONIO AUGUSTO FRANCO GARCIA**

Tese apresentada para obtenção do título de Doutor em Ciências. Área de concentração: Genética e Melhoramento de Plantas

Piracicaba
2012

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - ESALQ/USP

Mollinari, Marcelo
Desenvolvimento de um modelo para construção de mapas genéticos em autopoliploides, com aplicações em cana-de-açúcar / Marcelo Mollinari.- - Piracicaba, 2012.
98 p: il.

Tese (Doutorado) - - Escola Superior de Agricultura "Luiz de Queiroz", 2012.

1.Cana-de-açúcar 2. Mapeamento genético 3. Marcador molecular 4. Nucleotídeos 5. Polimorfismo 6. Poliploides I. Título

CDD 633.61
M726d

Aos meus pais, **Fani** e **Norberto**, dedico com todo meu amor.

Agradecimentos

Ao Professor Antonio Augusto Franco Garcia, que sabe qual é a verdadeira essência da palavra “orientar”. Augusto, é muito difícil descrever o quão grato sou por todos esses anos de ensinamentos e amizade. Muito Obrigado!

Ao Departamento de Genética da Escola Superior de Agricultura “Luiz de Queiroz” da Universidade de São Paulo, pela estrutura e pela oportunidade de realização deste doutorado.

À Professora Anete Pereira de Souza, pela confiança depositada no Laboratório de Genética Estatística da ESALQ/USP. Esse trabalho seria impossível de ser realizado sem essa valiosa colaboração.

Ao Professor Roland Vencovsky, por seus ensinamentos durante minha vida acadêmica.

Ao Dr. Thiago Marconi, que foi até a Austrália para aprender e trazer até nós essa impressionante ferramenta que são os marcadores SNPs usando espectrometria de massa.

À Dr[a]. Luciana Rosini e à Melina Mancini, por terem conduzido a população experimental e obtido os dados moleculares de microssatélites e AFLPs usados nesse trabalho.

Ao Dr. Oliver Serang, pela oportunidade de trabalharmos juntos no modelo para o “SNP call” em poliploides.

À FAPESP, pela concessão da bolsa e demais recursos ao longo desse trabalho (processo número: 2008/54402-4).

Aos amigos do Laboratório de Genética Estatística: Rodrigo, Maria Marta, Gabriel, Graciela, Renato, Edjane, João Ricardo, Carina, Luciano, Guilherme, Adriana, Rodrigo e Rafael, pelo companheirismo e discussões científicas ao longo de todos esses anos que ficamos juntos.

Aos professores do Departamento de Genética da Escola Superior de Agricultura “Luiz de Queiroz”.

Aos amigos e colegas do curso, pela convivência e aprendizado.

À minha irmã Flávia e ao meu cunhado Luis, pelo apoio incondicional.

À minha amada sobrinha Ana Clara, por existir na minha vida.

Aos funcionários do Departamento de Genética da ESALQ/USP: Seu Zé, Seu Antônio,

Valdir, Berdan, Léia, Macedônio e Fernandinho, pela convivência durante todos esses anos.

Ao Guilherme e ao João Ricardo, pela cuidadosa leitura dos originais e excelentes críticas e sugestões.

A todos amigos e familiares que direta ou indiretamente contribuíram para que esse trabalho fosse realizado. Muito Obrigado!

SUMÁRIO

RESUMO

Desenvolvimento de um modelo para construção de mapas genéticos em autopoliploides, com aplicações em cana-de-açúcar

Espécies autopoliploides são extremamente importantes na agricultura. No entanto, a estrutura complexa de seus genomas não é bem compreendida. Apesar de todos os avanços no mapeamento genético de autotetraploides, a grande maioria dos modelos utilizados para a construção de mapas em espécies autopoliploides com elevado nível de ploidia, tais como a cana-de-açúcar, são aproximações daqueles usados para organismos diplóides. Assim, este trabalho teve como objetivo o desenvolvimento de um novo modelo para construção de mapas genéticos em espécies autopoliploides com qualquer nível de ploidia e incluindo marcadores com todas as dosagens possíveis. Para tanto foi utilizada a tecnologia dos modelos de Markov ocultos. O modelo aqui apresentado pode ser aplicado a dados de marcadores dominantes e codominantes, com comportamento bialélico ou multialélico. O método baseia-se no cálculo das probabilidades condicionais que compõem a matriz de transição seguido da redução de sua dimensão usando uma abordagem computacional. O uso do método foi ilustrado em uma população de mapeamento de cana-de-açúcar proveniente do cruzamento entre duas variedades pré-comerciais (IACSP 95-3018 $\times$ IACSP 93-3046) e genotipadas com três tipos de marcadores: SNPs, microssatélites e AFLPs. Os resultados indicam que o novo método é muito eficiente na obtenção de mapas genéticos, mesmo em situações com níveis de ploidia altos e marcadores com altas doses, particularmente quando estes marcadores têm comportamento codominante. Também foi possível estimar a verossimilhança, as frações de recombinação e as fases de ligação usando a abordagem multiponto, a qual leva em consideração todos os marcadores do grupo de ligação analisado simultaneamente. O novo modelo aqui proposto representa um importante passo para realizar futuramente a localização de regiões genômicas associadas à variação das características quantitativas, no entendimento da arquitetura genética de tais características e na montagem de genomas de espécies autopoliploides.

Palavras-chave: Poliploides; SNP; Análise de ligação; Polimorfismo de nucleotídeo único

ABSTRACT

Development of a model to build genetic maps in autopolyploides, with applications in sugarcane

Autopolyploid species are extremely important in agriculture. However, the complex structure of their genomes is not well understood. Despite all advances in genetic mapping of autotetraploids, the vast majority of the models used for autopolyploid species with high ploidy level, such as sugarcane, are approximations of those used in diploid organisms. Thus, the aim of this work was to develop a new model to build genetic linkage maps in autopolyploid species with any ploidy level, including markers with all possible dosages. For doing so, hidden Markov model technology was used. The new model presented herein can be applied to dominant and codominant markers data, with biallelic or multiallelic behavior. The method is based on the calculation of conditional probabilities that comprise the transition matrix followed by a reduction of its dimension using a computer-based approach. An application of the method was illustrated with a sugarcane mapping population derived from a cross between two pre-commercial varieties (IACSP 95-3018 $\times$ IACSP 93-3046), scored with three types of markers: SNPs, microsatellite and AFLPs. The results indicate that the new method is very efficient in obtaining genetic maps, even for high ploidy levels and for markers with high dosages, particularly when these markers have codominant behavior. It was also possible to estimate the likelihood, the recombination fractions and the linkage phases between all markers using the multipoint approach, which takes into account all markers of the linkage group simultaneously. The new model proposed in this work represents a major step towards the location of genomic regions associated with variation in quantitative traits and its genetic architecture, and assembling autopoliploid genome sequences.

Keywords: Polyploids; SNP; Linkage analysis; Single nucleotide polymorphism

1 INTRODUÇÃO

O desenvolvimento de técnicas para obtenção de marcadores moleculares, combinado com eficientes métodos de análise dos dados, permitiu um melhor entendimento de vários fenômenos importantes para a genética e o melhoramento. Isto ocorre porque os marcadores fornecem informações a respeito da arquitetura genética dos caracteres quantitativos ao nível do DNA, tendo diversas aplicações tanto em estudos básicos como em pesquisas aplicadas (SOUZA, 2001). Inicialmente, os marcadores genéticos empregados baseavam-se em produtos da expressão gênica, sendo denominados marcadores morfológicos ou isoenzimáticos. Por serem de ocorrência rara e não estarem distribuídos de forma abundante ao longo do genoma, possuem uso limitado. Os marcadores moleculares, por sua vez, permitiram contornar esses problemas, sendo amplamente empregados. Dentre os marcadores moleculares mais usados atualmente, podem ser citados: RFLP (*Restriction Fragment Length Polymorphisms* - BOTSTEIN et al., 1980), AFLP (*Amplified Fragment Length Polymorphism* - VOS et al., 1995) e SSR (*Simple Sequence Repeat* - TAUTZ, 1989). Mais recentemente, surgiram os marcadores denominados SNP (*Single Nucleotide Polymorphism*), que têm grande potencial de aplicação em estudos de genética (SYVÄNEN, 2001). No caso particular das espécies poliploides, objeto de estudo da presente tese, o uso de SNPs é crucial para que avanços possam ser obtidos. Isto decorre não somente da grande abundância dos SNPs nos genomas, mas também da sua natureza codominante em espécies poliploides, tornando-os mais informativos e permitindo sensíveis avanços nos estudos genéticos.

A informação fornecida pelos marcadores moleculares pode ser útil de várias formas para o melhoramento genético, permitindo, por exemplo, a realização de estudos de divergência genética visando a predição de cruzamentos (LABORDA et al., 2005), a classificação dos indivíduos em grupos heteróticos (OLIVEIRA et al., 2004), a construção de mapas genéticos (MOLLINARI et al., 2009), o mapeamento de locos que controlam os caracteres quantitativos, ou QTLs (KAO; ZENG; TEASDALE, 1999; PASTINA et al., 2012) e até mesmo eventualmente a realização da chamada seleção assistida por marcadores (ZENG; KAO; BASTEN, 1999; MEUWISSEN; HAYES; GODDARD, 2001). Dentre essas aplicações, destaca-se o mapeamento de QTLs (*Quantitative Trait Loci*), por buscar um melhor entendimento da arquitetura genética justamente dos locos que controlam os caracteres quantitativos que são de difícil manuseio e, em geral, de maior importância para o melhoramento. A compreensão desses caracteres é muito importante

para que a seleção assistida seja implementada de forma efetiva, eventualmente com o desenvolvimento de novos métodos de melhoramento que incorporem as informações fornecidas pelos marcadores. Para que o mapeamento de QTLs seja realizado, contudo, é necessária a construção de mapas genéticos saturados, o que até o momento não foi feito de forma satisfatória para espécies poliploides, com exceção dos autotetraploides (LEACH et al., 2010). Os mapas também são importantes para realização de estudos evolutivos, de sintenia entre espécies, mapeamento associativo e até mesmo para montagem de genomas (LEWIN et al., 2009).

Hieter e Griffiths (1999) estimam que cerca de 50% das espécies vegetais são poliploides. Dentre as plantas de importância agronômica, podem ser citadas como exemplos de poliploides as culturas da cana-de-açúcar, algodão, banana, alfafa, batata, café e trigo, além de várias espécies forrageiras, como as do gênero *Brachiaria*. Os poliploides caracterizam-se pela presença de vários cromossomos por grupo de hom(e)ologia (GRIFFITHS et al., 2004), o que implica em mecanismos diferentes de pareamento na meiose, que pode ser preferencial ou aleatória (DOERGE; CRAIG, 2000). Como vantagens da poliploidia, podem ser citados a presença de mais de uma cópia de vários genes (redundância), o que eventualmente confere vantagens adaptativas (COMAI, 2005). Há na literatura vários relatos da presença de heterose como consequência do aumento do número de cópias do genoma (BIRCHLER; AUGER; RIDDLE, 2003; GRIFFITHS et al., 2004; COMAI, 2005; AUGER et al., 2005), sem contudo nenhuma explicação convincente sobre a base genética deste fenômeno. Como desvantagens, podem ser citadas as alterações na arquitetura celular e nos processos regulatórios, bem como dificuldades na meiose (COMAI, 2005).

Estudos sobre expressão gênica em *Saccharomyces cerevisiae* mostraram que diferentes níveis de ploidia foram responsáveis por diferentes níveis de expressão gênica (GALITSKI et al., 1999); em outras palavras, parece haver associação entre a dose dos alelos e sua expressão. O entendimento de tais processos permitiria definir estratégias de seleção assistida específicas para poliploides, uma vez que tal mecanismo não é perfeitamente entendido nem usado de forma consciente nos processos de seleção. Há também relatos de que a regulação gênica também dependa do nível de ploidia (HIETER; GRIFFITHS, 1999; GUO; DAVIS; BIRCHLER, 1996; OSBORN et al., 2003). Soltis e Soltis (1999) mencionam que várias questões relevantes poderiam ser respondidas se houvesse mais estudos estudos genéticos sobre poliploides.

Apesar dos avanços obtidos nos estudos de mapeamento genético em espécies com disponibilidade de linhagens endogâmicas (como os retrocruzamentos e as populações F_2), a aplicação de tais técnicas e métodos de análise genético-estatística em espécies poliploides ainda

é bastante restrita (DOERGE; CRAIG, 2000; GAZAFFI, 2009; PASTINA et al., 2010; GAZAFFI et al., 2010). Isto ocorre porque várias dificuldades são encontradas para se realizar mapeamento genético nesta situação, incluindo aumento considerável no número de genótipos presentes nas populações, dificuldades em determinar o número de cópias (dose) de cada loco do marcador e dos QTLs, dificuldades em observar os eventos de recombinação em função da presença de cópias adicionais, e alterações no pareamento cromossômico na meiose (COMAI, 2005; DOERGE; CRAIG, 2000). Quando são incluídos SNPs, novos desafios também surgem para a análise dos dados. Em primeiro lugar, os programas computacionais que são fornecidos com os equipamentos de genotipagem comumente usados (como o software BeadStudio, distribuído juntamente com a plataforma Illumina GoldenGate™ e o software TYPER, distribuído juntamente com a plataforma Sequenom iPLEX MassARRAY®) não podem ser utilizados diretamente para classificar os dados nestas espécies (dito *"SNP genotype call"*). A utilização desses resultados para construção de mapas genéticos também não é possível até o presente momento, com exceção das espécies autotetraploides, para as quais já existem métodos adequados de análise à disposição (LEACH et al., 2010).

A cana-de-açúcar pode ser usada como modelo para o desenvolvimento de métodos de análise genético-estatística que permitiriam a realização de estudos avançados em autopoliploides. Isto ocorre em função da complexa organização genômica desta espécie. As variedades modernas surgiram de cruzamentos entre genótipos resultantes de cruzamentos interespecíficos principalmente entre *Saccharum officinarum* (variedades domesticadas, geralmente com $2n = 8x = 80$) e *S. spontaneum*, que é uma espécie selvagem (com $2n = 64, 80, 96, 112$ ou 128, e $x = 8$) (PIPERIDIS; PIPERIDIS; D'HONT, 2010). É ainda comum a presença de aneuploidia, ou seja, número de cromossomos variável em cada grupo de hom(e)ologia (GRIVET; ARRUDA, 2001). Novos métodos para construção de mapas genéticos que venham a ser desenvolvidos precisam incorporar tais características. Dado que tal complexidade dificilmente será encontrada em outras espécies, possivelmente os novos modelos poderão ser usados para vários outros poliploides ainda pouco estudados.

A principal abordagem usada para construção de mapas genéticos em cana-de-açúcar baseia-se no método desenvolvido por Wu et al. (1992), posteriormente expandido por Da Silva e Sorrells (1996) e Ripol et al. (1999). Percebendo que, independente do nível de ploidia da espécie, marcadores em dose única sempre segregam na proporção 1:1 num população segregante resultante do cruzamento de dois genitores não endogâmicos, estes autores propuseram o uso de marcadores com essa dose para construção de mapas de cana-de-açúcar, o que foi também se-

guido em outros poliploides. Sua vantagem é possibilitar o uso de programas de computador, como por exemplo o MAPMAKER/EXP, (LANDER et al., 1987), já disponíveis para outras espécies, uma vez que tal padrão de segregação é o mesmo observado em retrocruzamentos. O fato da fase de ligação entre os marcadores ser desconhecida foi resolvido com emprego da abordagem conhecida como *duplo pseudo-testcross* (GRATTAPAGLIA; SEDEROFF, 1994), que resulta na construção de um mapa genético para cada um dos genitores.

Há várias limitações nessas abordagens que precisam ser resolvidas para que avanços sejam possíveis no estudo genético dos poliploides. Em primeiro lugar, usar apenas locos que possuam um determinado tipo de segregação (1:1, dose única) para construir os mapas genéticos pode resultar em mapas baseados em apenas uma pequena parte do genoma, já que não há nenhuma razão biológica para assumir que locos em dose única sejam maioria no genoma. Em partes posteriores desta tese será demonstrado que as estimativas dos locos em dose única no genoma são fortemente enviesadas e não devem representar o que de fato ocorre. Sobre os *duplo pseudo-testcross*, vale mencionar que essa estratégia fornece dois mapas de ligação, sendo que em cada um deles há marcadores segregando exclusivamente para um determinado genitor, uma vez que o outro genitor apresenta alelos fixados para os mesmos locos (GRATTAPAGLIA; SEDEROFF, 1994). A integração das informações contidas nesses mapas individuais em um único mapa somente pode ser feita com a presença de marcadores em heterozigose em ambos os parentais, os quais são utilizados para estabelecer relações de ligação entre os marcadores segregando individualmente em cada genitor (WU et al., 2002b; GARCIA et al., 2006; OLIVEIRA et al., 2007, 2008). Não é correto assumir que a maioria dos locos do genoma segrega em apenas um dos genitores, especialmente em poliploides, e dados experimentais que serão apresentados nesta tese corroboram tal afirmativa. A construção de um mapa genético integrado, utilizando-se marcadores com diferentes tipos de segregação simultaneamente, além de mais realista do ponto de vista biológico, apresenta inúmeras vantagens, pois permite aumentar a saturação do mapa de ligação e estender a caracterização da variação polimórfica em todo o genoma.

Neste contexto, a presente tese teve como objetivo o desenvolvimento de um novo método genético-estatístico para construção de mapas genéticos multiponto em espécies autopoliploides, incorporando todas as doses possíveis e diferentes combinações de polimorfismo nos genitores, resultando em mapas integrados. Foi dada ênfase em dados de marcadores do tipo SNP, que possuem características únicas e muito vantajosas para esse tipo de análise. Porém, é importante mencionar que locos obtidos com outros tipos de marcadores (como por exemplo os AFLPs e microssatélites) também podem ser incorporados aos mapas genéticos que forem construídos

com o método aqui proposto.

2 REVISÃO BIBLIOGRÁFICA

Neste capítulo, serão apresentados os principais conceitos e métodos usados para a construção de mapas genéticos. Num primeiro momento, serão detalhados os métodos clássicos apresentados para espécies diploides com linhagens endogâmicas à disposição. Isto é necessário para fundamentar as ideias de testes de ligação, obtenção de estimativas multiponto e ordenação dos marcadores. Em seguida, será abordado o principal método usado em espécies diploides que não dispõem de linhagens, mas que vem sendo usado em cana-de-açúcar quando se deseja construir mapas integrados. Será ainda apresentado o método utilizado para inclusão de marcadores com outras doses baseado em teste de dois pontos. Finalmente, o principal método desenvolvido para espécies autotetraploides será detalhado, e será discutido para demonstrar que tal abordagem não pode ser usada para outras espécies com maiores níveis de ploidia. O novo método aqui proposto será detalhado no item Material e Métodos.

2.1 Construção de Mapas Genéticos

Os trabalhos de Mollinari (2007) e Mollinari et al. (2009) apresentam em detalhes os conceitos e procedimentos usados para construção de mapas genéticos em espécies diploides que permitem a obtenção de linhagens homozigóticas, e que portanto têm à disposição populações de retrocruzamentos, F_2 e linhagens recombinantes. Ambos foram usados como base para a redação do presente tópico. Surpreendentemente, há na literatura poucos textos que apresentem tais conceitos de maneira linear, ou seja, num mesmo contexto que indique sequencialmente como os mapas devem ser construídos. Como leitura adicional para os conceitos estatísticos, pode ser recomendado o livro de Liu (1998), ou mesmo o tutorial dos programas computacionais R/QTL (BROMAN, 2010) e ONEMAP (MARGARIDO; MOLLINARI; GARCIA, 2011).

2.1.1 Fundamentos

De forma geral, a construção de mapas de ligação é feita em duas etapas: i) agrupamento dos marcadores moleculares nos chamados *grupos de ligação*; ii) estimação da ordem e da distância entre os marcadores dentro destes grupos. É necessário portanto que indivíduos de uma

população segregante sejam genotipados. No caso da etapa i, os principais testes usados são o teste de qui-quadrado (χ^2) e o teste da razão de verossimilhança (MOLLINARI, 2007). A etapa ii é geralmente mais trabalhosa e diversos algoritmos e critérios estão disponíveis, sendo que as estimativas das distâncias geralmente empregam métodos multiponto, baseados em cadeias de Markov ocultas (MOLLINARI et al., 2009).

A escolha da melhor ordem faz parte de uma classe de problemas matemáticos conhecidos como "classe dos problemas de tempo polinomial não determinístico de difícil resolução" (*NP Hard Problem*) (MESTER et al., 2003), conhecidos popularmente como problema do caixeiro viajante (*TSP - Traveling Salesman Problem*). Neste tipo de problema, é impossível obter uma solução exata única, então métodos aproximados (chamados de "heurísticas" em ciência da computação) são necessários. Mollinari et al. (2009), usando simulações, estudaram essa classe de problemas e apresentaram recomendações gerais sobre como ordenar os marcadores em diversas situações encontradas pelos geneticistas atualmente.

Uma vez ordenados os marcadores, são estimadas as distâncias entre os mesmos. Para transformar as frações de recombinação (que são probabilidades de ocorrência de recombinação) em medidas aditivas, são comumente empregadas as chamadas funções de mapeamento (LIU, 1998). As principais são:

$$
\begin{aligned}
d_{ij} &= -\frac{1}{2}\ln(1-2r_{ij}) \quad \text{(HALDANE, 1919)} \\
d_{ij} &= r_{ij} \quad \text{(MORGAN, 1928)} \\
d_{ij} &= \frac{1}{4}\ln\left[\frac{1+2r_{ij}}{1-2r_{ij}}\right] \quad \text{(KOSAMBI, 1944)}
\end{aligned}
$$

sendo r_{ij} a fração de recombinação entre os locos i e j, e d_{ij} a distância (em Morgans) entre os locos i e j.

2.1.2 Cadeias de Markov

Uma categoria de modelos de fulcral importância para a construção de mapas genéticos são as chamadas cadeias de Markov. Por permitirem que toda a informação dos grupos de ligação seja utilizada simultaneamente no momento da estimação das distâncias entre locos e no cálculo de verossimilhança de uma dada ordem, devem ser empregadas sempre que possível (MOLLINARI et al., 2009). Cadeias de Markov são modelos matemáticos usados para estudar

processos resultantes de sequências de observações, exatamente como ocorre nos mapas genéticos (MOLLINARI, 2007). Sua utilização na genética foi sugerida inicialmente no artigo seminal de Lander e Green (1987). Vale mencionar que, no caso dos mapas genéticos, uma sub-categoria de processos relacionados à cadeia, também conhecida como processo markoviano oculto, deve ser usada e por essa razão será aqui apresentada com base em Mollinari (2007). Para maiores detalhes, consultar (RABINER, 1989), (BROMAN; SEN, 2009) e (JIANG; ZENG, 1997).

Para introduzir o assunto, Mollinari (2007) apresenta um exemplo simples, aqui incluído. Considere que o clima de um determinado local pode apresentar um dos três estados: *chuvoso*, *nublado* ou *ensolarado*. Considere agora uma sequência de quatro dias e os respectivos estados que o clima pode apresentar nesses dias, como por exemplo, *chuvoso*, *nublado*, *chuvoso* e *ensolarado*. O estado *chuvoso* pode ser representado por E_1, o estado *nublado* por E_2 e o estado *ensolarado* por E_3; os dias podem ser representados por $t = 1, 2, 3, 4$ e o número de estados por $N = 3$. Então, outra forma de representar a sequência de observações exemplificada seria $\{E_1, E_2, E_1, E_3\}$, correspondendo aos dias $t = 1, 2, 3, 4$. Deseja-se calcular a probabilidade de ocorrência desta sequência de observações.

O modelo de Markov assume *independência condicional* entre as observações, ou seja, que a probabilidade de ocorrência de uma determinada condição climática em um dia depende exclusivamente da condição climática do dia anterior. Define-se como *probabilidade de transição* a probabilidade de observar o estado E_j no tempo t, dado que no tempo $t - 1$ o estado observado foi E_i. Usualmente, estas probabilidades são organizadas matricialmente, sendo que o número de linhas e o número de colunas é igual ao número de estados N. Esta matriz quadrada é chamada de *matriz de transição*, e pode ser denotada por

$$\mathbf{H} = [h_{ij}]_{N \times N}$$

sendo h_{ij} a probabilidade do estado atual ser E_j dado que o estado anterior foi E_i.

A probabilidade da sequência de observações $\{E_1, E_2, E_1, E_3\}$ pode ser calculada multiplicando-se as probabilidades de transição entre os estados da sequência:

$$P[\{E_1, E_2, E_1, E_3\}|Modelo] = h_{12} \times h_{21} \times h_{13}$$

Porém, em muitos fenômenos importantes, os estados entre as transições da cadeia não são observados. (RABINER, 1989) apresenta um exemplo simples usando diferentes urnas contendo

diferentes proporções de bolas coloridas. Se apenas uma sequência de bolas foi observada, é impossível saber com certeza de que urnas tais bolas são provenientes. No entanto, é possível incluir no modelo a chamada *função de emissão*, que associa cada cor de bola com uma dada urna. Por exemplo, se uma das urnas contiver apenas bolas de uma dada cor, caso uma bola de outra cor seja observada, sabe-se com certeza que ela não é proveniente desta urna. No entanto, se essa cor for observada, é possível calcular a probabilidade de ela ter vindo de cada uma das urnas.

No caso dos mapas genéticos, o modelo de cadeia de Markov oculta permite modelar o fenômeno de maneira bastante realista. Quando os indivíduos são genotipados com algum marcador, busca-se conhecer o real genótipo dos indivíduos para o loco em questão. Se o marcador for completamente informativo (codominante, no caso de diploides) e se não houver nenhum erro de genotipagem, a correspondência é direta. Por exemplo, se o genótipo do marcador for heterozigoto, pode-se afirmar que o genótipo do indivíduo é heterozigoto naquele loco. A genotipagem de uma série de locos ligados e ordenados caracterizam a cadeia de Markov. Porém, em situações práticas, ocorrem perdas de dados para alguns locos, e portanto não é possível ao certo saber o genótipo do indivíduo para aquele loco, e faz mais sentido assumir então que a cadeia tem estados ocultos. No entanto, é possível calcular a probabilidade do indivíduo ter um dado genótipo para o loco com base em todo o grupo de ligação simultaneamente. Estados ocultos ocorrem também quando há erros de genotipagem (o que é observado pelo marcador não corresponder ao real genótipo do loco), ou ainda marcadores dominantes (em que alguns genótipos são possíveis para uma dada observação para o marcador).

Para melhor compreensão deste modelo, será apresentado novamente um exemplo didático, extraído de (MOLLINARI, 2007). Considere que existam quatro urnas numeradas. Dentro de cada urna existe uma proporção diferente de quatro cores de bolas: *vermelha*, *azul*, *amarela* e *verde*. Imagine que o seguinte experimento seja realizado: escolhe-se uma urna aleatoriamente, retira-se uma bola e anota-se a cor. A bola é então reposta na urna de origem. Em seguida, escolhe-se outra urna, que pode inclusive ser a mesma, de acordo com uma probabilidade de transição, e repete-se o processo. Procede-se novamente dessa forma por 10 vezes. Neste caso, é de interesse saber qual é a probabilidade de observar uma determinada sequência de cores, sem saber de quais urnas foram retiradas as bolas. Nota-se que o experimento pode ser modelado usando a cadeia de Markov oculta.

Neste caso, os estados são denotados por $E = \{E_1, E_2, \ldots, E_N\}$ (N estados), os símbolos das observações são denotados por $V = \{v_1, v_2, \ldots v_U\}$ (U símbolos), o estado atual é denotado

por τ_t, o número de observações é denotado por Y e a sequência de observações é denotada por $O = O_1, O_2, \ldots, O_Y$. No exemplo, $E = \{E_1, E_2, E_3, E_4\}$ (urnas), $V = \{v_1, v_2, v_3, v_4\}$ (vermelho, azul, verde e amarelo) e $Y = 10$. Verificando-se a proporção das cores das bolas nas urnas, é possível inferir sobre a urna de origem de cada bola, associando alguma probabilidade a essa inferência.

A função de probabilidade (ou de emissão) que liga os estados (urnas) às observações (cores) é definida por:

$$b_j(w) = P[v_w \text{ em } t | \tau_t = E_j], \qquad 1 \leq j \leq N \text{ e } 1 \leq w \leq U.$$

A distribuição de probabilidades dos símbolos das observações para um determinado estado E_j é

$$\mathbf{B} = \{b_j(w)\}$$

Esta distribuição fornece a probabilidade de observar um determinado símbolo dado um determinado estado. Pode ainda ser definida a distribuição de probabilidade inicial, que indica a probabilidade do estado inicial τ_1 ser E_i:

$$\pi = \{\pi_i\}, \text{sendo} \qquad \pi_i = P[\tau_1 = E_i], \qquad 1 \leq i \leq N.$$

Por conveniência, a matriz de transição $\mathbf{H}$, a distribuição de probabilidade $\mathbf{B}$ e a distribuição de probabilidade inicial π podem ser representadas por λ (parâmetros do modelo):

$$\lambda = (\mathbf{H}, \mathbf{B}, \pi)$$

Finalmente, a probabilidade da sequência de observações O dado o modelo λ, pode ser obtida através da seguinte equação:

$$P(O|\lambda) = \sum_{\tau_1, \tau_2, \ldots, \tau_T} \pi_{\tau_1} b_{\tau_1}(O_1) h_{\tau_1 \tau_2} b_{\tau_2}(O_2) \ldots h_{\tau_{T-1} \tau_T} b_{\tau_T}(O_T)$$

2.1.3 Verossimilhança Usando Cadeias de Markov Ocultas no Contexto do Mapeamento Genético

Os principais critérios usados para comparar diferentes ordens dos locos dentro dos gru-

pos de ligação são verossimilhança multiponto, usando cadeias de Markov ocultas (LANDER; GREEN, 1987), SARF (*sum of adjacent recombination fractions* - FALK, 1989), PARF (*product of adjacent recombination fractions* - WILSON, 1988), e SALOD (*sum of adjacent lod scores* - WEEKS; LANGE, 1987). Usando extensas simulações computacionais, (MOLLINARI, 2007) e (MOLLINARI et al., 2009) discutem em que situações cada um desses critérios é mais adequado. Como no presente caso foi utilizado o primeiro deles, apenas ele será aqui detalhado. As razões que motivaram tal escolha será discutida no item Material e Métodos.

O texto a seguir é uma adaptação do que foi apresentado por (JIANG; ZENG, 1997) e (MOLLINARI, 2007). Considere como exemplo uma população F_2 proveniente do cruzamento de duas linhagens homozigotas e diploides P_1 e P_2. Suponha que existam m marcadores posicionados em um mapa com ordem conhecida $M_1, \ldots, M_m$. O *genótipo* do marcador $M_k, k = 1, \ldots, m$, será denotado como x_k e pode assumir valores 1, 0 e -1 se M_k for homozigoto como P_1, heterozigoto ou homozigoto como P_2, respectivamente. Seja o *fenótipo* do marcador M_k, para o mesmo indivíduo, denotado por z_k. Se o genótipo do marcador é completamente informativo (codominante e sem erro de genotipagem), o fenótipo será igual ao genótipo: $z_k = x_k = \{1\}$, $z_k = x_k = \{0\}$ou $z_k = x_k = \{-1\}$. Em outras situações, como quando há um dado perdido, o genótipo é desconhecido e $z_k = \{1, 0, -1\} = M$ (*missing*). Quando o marcador é parcialmente informativo (marcador dominante), pode haver mais de um genótipo associado ao fenótipo observado.

A sequência $M_1, \ldots, M_m$, conforme já mencionado, pode ser modelada como uma cadeia de Markov oculta. Vale lembrar que a propriedade markoviana de independência condicional pode ser assumida, uma vez que violações nesta suposição serão corrigidas com emprego de alguma função de mapeamento. Em outras palavras, a propriedade markoviana assume ausência de interferência, o que pode não ocorrer em fenômenos biológicos. Porém, a função de mapeamento mais usada (Kosambi) corrige eventuais distorções nesta pressuposição. Para detalhes, ver (LIU, 1998).

Neste contexto, os estados ocultos são os genótipos dos marcadores e os estados observáveis são os fenótipos. Caso todos os marcadores sejam codominantes e não existam dados perdidos, a cadeia não será oculta e a função de ligação será a identidade. A matriz de transição entre os genótipos dos locos é composta por elementos que são função da fração de recombinação entre eles:

$$\begin{aligned}\mathbf{H}(r_k) &= \begin{bmatrix} P(x_{k+1}=1|x_k=1) & P(x_{k+1}=0|x_k=1) & P(x_{k+1}=-1|x_k=1) \\ P(x_{k+1}=1|x_k=0) & P(x_{k+1}=0|x_k=0) & P(x_{k+1}=-1|x_k=0) \\ P(x_{k+1}=1|x_k=-1) & P(x_{k+1}=0|x_k=-1) & P(x_{k+1}=-1|x_k=-1) \end{bmatrix} \\ &= \begin{bmatrix} (1-r_k)^2 & 2r_k(1-r_k) & r_k^2 \\ r_k(1-r_k) & (1-r_k)^2+r_k^2 & r_k(1-r_k) \\ r_k^2 & 2r_k(1-r_k) & (1-r_k)^2 \end{bmatrix}\end{aligned}$$

sendo r_k a fração de recombinação entre os marcadores M_k e M_{k+1}.

As funções de emissão dependem do tipo de genótipo x_k dos marcadores, que num F_2 podem ser denotados como 1, 0 ou -1. Por exemplo, para $x_k = 1$ (i.e., homozigoto dominante), a distribuição de probabilidades de emissão considerando-se um marcador dominante é

$$\mathbf{B} = \{b_1(z_k)\} = \begin{cases} 1 & \text{se } z_k = D \\ 0 & \text{se } z_k = -1 \end{cases}$$

e assim sucessivamente para os demais genótipos e tipos de marcador. A distribuição de probabilidades inicial (*a priori*) é simplesmente função das proporções mendelianas para o tipo de população experimental considerada. No caso:

$$\pi = \{\pi_{x_k}\} = \begin{cases} 1/4 & \text{se } x_k = 1 \\ 1/2 & \text{se } x_k = 0 \\ 1/4 & \text{se } x_k = -1 \end{cases}$$

Usando a notação matricial de (JIANG; ZENG, 1997) para cálculo da verossimilhança de uma dado grupo de ligação, tais probabilidades são indicadas no seguinte vetor:

$$\mathbf{q}' = \begin{bmatrix} \frac{1}{4}, & \frac{1}{2}, & \frac{1}{4} \end{bmatrix}$$

São então definidas matrizes indicadoras $\mathbf{I}_{z_j}$, sendo $z_j = M = \{-1, 0, 1\}$ (*missing*), $D = \{1, 0\}$ (*dominant*), $R = \{0, 1\}$ (*recessive*), 1, 0 ou -1, que dependem do conteúdo de informação do fenótipo do marcador M_j). Tais matrizes são usadas para pré-multiplicar a ma-

triz de transição, eliminando genótipos que não podem ocorrer dado o fenótipo observado:

$$\mathbf{I}_M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{I}_D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{I}_R = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{I}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{I}_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{I}_{-1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Quando o fenótipo do marcador é D, por exemplo, $\mathbf{I}_D$ elimina a última linha da matriz de transição na multiplicação $\mathbf{I}_D\mathbf{H}(r_k)$, uma vez que o fenótipo $z_k = D$ exclui a possibilidade do genótipo ser $x_k = -1$.

Finalmente, para cálculo da verossimilhança do grupo de ligação, sejam $z_{l1}, \ldots, z_{lm}$ as observações dos fenótipos dos marcadores $M_1, \ldots, M_m$ para o l-ésimo indivíduo em uma população F_2. A verossimilhança é dada por

$$L = \prod_{l=1}^{n} \mathbf{q}'\mathbf{I}_{z_{l1}}\mathbf{H}(r_1)\mathbf{I}_{z_{l2}}\mathbf{H}(r_2)\mathbf{I}_{z_{l3}}, \ldots, \mathbf{H}(r_{m-1})\mathbf{I}_{z_{lm}}\mathbf{c}$$

sendo n o número de indivíduos e $\mathbf{c} = \{1\}_{3\times 1}$. Esta verossimilhança pode ser usada como critério para a comparação de diferentes ordens.

As frações de recombinação entre todos os marcadores podem ser estimadas através do algoritmo EM (DEMPSTER; LAID; RUBIN, 1977), num processo iterativo. Como os cálculos envolvem todos os marcadores simultaneamente, essas estimativas são ditas multiponto, e evidentemente possuem mais acurácia que aquelas obtidas apenas com análises de dois pontos.

2.1.4 Mapeamento Genético em Irmãos Completos

Até o presente momento, a grande maioria dos mapas genéticos de cana-de-açúcar tem sido construídos com modelos desenvolvidos para populações de espécies diploides derivadas de linhagens endogâmicas, a saber, os retrocruzamentos. Entretanto, há grandes vantagens em se utilizar expansões deste conceito para cana-de-açúcar, mesmo que apenas locos em dose única sejam considerados. Surgem neste cenário os métodos desenvolvidos para espécies que não possuem linhagens endogâmicas e que vem sendo usados com sucesso (GARCIA et al., 2006;

OLIVEIRA et al., 2007; PASTINA et al., 2012) e está implementado na última versão (2.0-1) do software ONEMAP (MARGARIDO; SOUZA; GARCIA, 2007; MARGARIDO; MOLLINARI; GARCIA, 2011). O algoritmo multiponto descrito a seguir (seção 2.1.4.1) foi desenvolvido utilizando-se ideias propostas por Jiang e Zeng (1997) , Wu et al. (2002a) e Wu et al. (2002b). Como tais ideias foram apresentadas para organismos diploides, o algoritmo será apresentado para diversos padrões de segregação, incluindo aqueles comumente encontrados em marcadores em dose simples em cana-de-açúcar, ou seja, com segregação 1:1 e 3:1, sendo que estes últimos permitem a integração do mapa genético.

2.1.4.1 Algoritmo para Cálculo da Verossimilhança e Reconstrução do Mapa Integrado Usando a Cadeia de Markov Oculta

Considere uma família de irmãos completos derivada do cruzamento de dois indivíduos diploides não homozigóticos P e Q. Sejam $P_k^{\{1,2\}}$ e $P_{k+1}^{\{1,2\}}$, e $Q_k^{\{1,2\}}$ e $Q_{k+1}^{\{1,2\}}$ dois locos adjacentes em P e Q, respectivamente. Os sobrescritos $\{1,2\}$ em P e Q indicam o conjunto de possíveis alelos nos locos e cada loco pode ter até dois alelos diferentes em cada genitor. w denota as quatro possíveis configurações das fases de ligação entre estes dois locos (Figura 1). A configuração $w = 1$ denota fase de associação entre os alelos nos dois genitores, $w = 2$ e $w = 3$ denotam fase de repulsão nos genitores Q e P, respectivamente e $w = 4$ denota fase de repulsão em ambos genitores.

Wu et al. (2002a) apresentam uma tabela mostrando todas as possíveis combinações de genótipos dos marcadores, seu cruzamento e os padrões de banda observados nos genitores e em sua progênie resultante desses cruzamentos. Para cada loco existem até quatro alelos além de um alelo nulo. Uma reprodução parcial da Tabela 1 de (WU et al., 2002a) pode ser vista na Tabela 1. Foram classificados 18 tipos de cruzamento e os quatro alelos foram denotados pelas letras a, b, c e d; o alelo nulo foi representado pela letra o. Os alelos a, b, c e d são codominantes entre si e dominantes em relação a o. Estes alelos podem ser usados para identificar o tipo de cruzamento trocando os alelos $\{1,2\}$, mostrados acima, pela letra correspondente. Por exemplo, se um marcador é A e o tipo de cruzamento é 1 (cruzamento $ab \times cd$), $P_k^1 = P_k^a$, $P_k^2 = P_k^b$, $Q_k^1 = Q_k^c$ e $Q_k^2 = Q_k^d$.

Para reconstruir um mapa genético de populações de irmãos completos provenientes de genitores heterozigóticos usando cadeias de Markov ocultas, é necessário que sejam levados em conta os vários tipos de cruzamentos apresentados na Tabela 1 e fases da ligação entre os

$$
\begin{array}{lcccc}
 & P & & Q & \\
w_k = 1\text{:} & \begin{array}{c|c} P_k^1 & P_k^2 \\ P_{k+1}^1 & P_{k+1}^2 \end{array} & \times & \begin{array}{c|c} Q_k^1 & Q_k^2 \\ Q_{k+1}^1 & Q_{k+1}^2 \end{array} & \\
 & \text{or} & & & \\
w_k = 2\text{:} & \begin{array}{c|c} P_k^1 & P_k^2 \\ P_{k+1}^1 & P_{k+1}^2 \end{array} & \times & \begin{array}{c|c} Q_k^1 & Q_k^2 \\ Q_{k+1}^2 & Q_{k+1}^1 \end{array} & \\
 & \text{or} & & & \\
w_k = 3\text{:} & \begin{array}{c|c} P_k^1 & P_k^2 \\ P_{k+1}^2 & P_{k+1}^1 \end{array} & \times & \begin{array}{c|c} Q_k^1 & Q_k^2 \\ Q_{k+1}^1 & Q_{k+1}^2 \end{array} & \\
 & \text{or} & & & \\
w_k = 4\text{:} & \begin{array}{c|c} P_k^1 & P_k^2 \\ P_{k+1}^2 & P_{k+1}^1 \end{array} & \times & \begin{array}{c|c} Q_k^1 & Q_k^2 \\ Q_{k+1}^2 & Q_{k+1}^1 \end{array} &
\end{array}
$$

Figura 1 – Quatro possíveis fases de ligação para os marcadores nos cromossomos dos genitores

marcadores nos cromossomos dos genitores (Figura 1).

Seja $M_1, \cdots, M_m$ um conjunto de m marcadores ordenados em um cromossomo e $w_1, \cdots, w_{m-1}$ as fases de ligação entre estes marcadores nos cromossomos dos genitores. $x_{j,k} = \{P_k^1 Q_k^1\}, \{P_k^1 Q_k^2\}, \{P_k^2 Q_k^1\}, \{P_k^2 Q_k^2\}$ denota o genótipo dos marcadores M_k para o j-ésimo indivíduo na progênie. De maneira similar, $z_{j,k}$ denota o fenótipo de M_k para o mesmo indivíduo. Dependendo do tipo do marcador M_k, pode existir uma mistura de classes genotípicas e $z_{j,k}$ pode assumir diferentes valores:

$$z_{j,k} = \begin{cases} \{P_k^1Q_k^1\},\{P_k^1Q_k^2\},\{P_k^2Q_k^1\},\{P_k^2Q_k^2\}, & \text{se } M_k \text{ é do tipo A} \\ \{P_k^1Q_k^1,P_k^1Q_k^2\},\{P_k^2Q_k^1\},\{P_k^2Q_k^2\}, & \text{se } M_k \text{ é do tipo B}_1 \\ \{P_k^1Q_k^1,P_k^2Q_k^1\},\{P_k^1Q_k^2\},\{P_k^2Q_k^2\}, & \text{se } M_k \text{ é do tipo B}_2 \\ \{P_k^1Q_k^2,P_k^2Q_k^1\},\{P_k^1Q_k^1\},\{P_k^2Q_k^2\}, & \text{se } M_k \text{ é do tipo B}_3 \\ \{P_k^1Q_k^1,P_k^1Q_k^2,P_k^2Q_k^1\},\{P_k^2Q_k^2\}, & \text{se } M_k \text{ é do tipo C} \\ \{P_k^1Q_k^1,P_k^1Q_k^2\},\{P_k^2Q_k^1,P_k^2Q_k^2\}, & \text{se } M_k \text{ é do tipo D}_1 \\ \{P_k^1Q_k^1,P_k^2Q_k^1\},\{P_k^1Q_k^2,P_k^2Q_k^2\}, & \text{se } M_k \text{ é do tipo D}_2 \\ \{P_k^1Q_k^1,P_k^1Q_k^2,P_k^2Q_k^1,P_k^2Q_k^2\}, & \text{se } M_k \text{ é dado perdido} \end{cases}$$

Dois ou mais valores entre chaves indicam os possíveis genótipos para um dado fenótipo molecular. Por exemplo, marcadores do tipo A são completamente informativos e não causam estados ocultos enquanto que marcadores do tipo D_1 tipo de cruzamento 9 ($ab \times cc$) causam estados ocultos em $\{P_k^1Q_k^1, P_k^1Q_k^2\}$ e $\{P_k^2Q_k^1, P_k^2Q_k^2\}$, uma vez que é impossível identificar a origem dos alelos Q_k^1 e Q_k^2.

Seja uma sequência de genótipos $x_{j,1}, \cdots, x_{j,m}$ e sua correspondente sequência de fenótipos $z_{j,1}, \cdots, z_{j,m}$ para o j-ésimo indivíduo na progênie. Estas sequências podem ser vistas como sequências de estados ocultos e sequências de observações, respectivamente, uma vez

Tabela 1 – Possíveis combinações de genótipos dos marcadores, seu cruzamento e os padrões de banda observados nos genitores e em sua progênie

			Genitores		Progênie	
		Tipo de cruzamento	Cruzamento	Bandas observadas	Bandas observadas	Segregação
A		1	$ab \times cd$	$ab \times cd$	ac, ad, bc, bd	1:1:1:1
		2	$ab \times ac$	$ab \times ac$	a, ac, ba, bc	1:1:1:1
		3	$ab \times co$	$ab \times c$	ac, a, bc, b	1:1:1:1
		4	$ao \times bo$	$a \times b$	ab, a, b, o	1:1:1:1
B	B_1	5	$ab \times ao$	$ab \times a$	$ab, 2a, b$	1:2:1
	B_2	6	$ao \times ab$	$a \times ab$	$ab, 2a, b$	1:2:1
	B_3	7	$ab \times ab$	$ab \times ab$	$a, 2ab, b$	1:2:1
C		8	$ao \times ao$	$a \times a$	$3a, o$	3:1
D	D_1	9	$ab \times cc$	$ab \times c$	ac, bc	1:1
		10	$ab \times aa$	$ab \times a$	a, ab	1:1
		11	$ab \times oo$	$ab \times o$	a, b	1:1
		12	$bo \times aa$	$b \times a$	ab, a	1:1
		13	$ao \times oo$	$a \times o$	a, o	1:1
	D_2	14	$cc \times ab$	$c \times ab$	ac, bc	1:1
		15	$aa \times ab$	$a \times ab$	a, ab	1:1
		16	$oo \times ab$	$o \times ab$	a, b	1:1
		17	$aa \times bo$	$a \times b$	ab, a	1:1
		18	$oo \times ao$	$o \times a$	a, o	1:1

que a última pode ser diretamente acessada através do padrão das bandas. A probabilidade $P\left(P_{k+1}^{\{1,2\}}Q_{k+1}^{\{1,2\}} \mid P_k^{\{1,2\}}Q_k^{\{1,2\}}\right)$ de um genótipo particular ocupar a posição $k+1$ dado um certo genótipo na posição k é chamada de probabilidade de transição e depende da fração de recombinação r e da fase de ligação w entre os locos k e $k+1$. Para $w=1$, a matriz de probabilidades de transição $\mathbf{H}_{k,r}^{w}$ pode ser expressa como (WU et al., 2002a)

$$\begin{aligned}
\mathbf{H}_{k,r}^{1} &= \begin{bmatrix}
P\left(P_{k+1}^{1}Q_{k+1}^{1} \mid P_k^{1}Q_k^{1}\right) & P\left(P_{k+1}^{1}Q_{k+1}^{2} \mid P_k^{1}Q_k^{1}\right) & P\left(P_{k+1}^{2}Q_{k+1}^{1} \mid P_k^{1}Q_k^{1}\right) & P\left(P_{k+1}^{2}Q_{k+1}^{2} \mid P_k^{1}Q_k^{1}\right) \\
P\left(P_{k+1}^{1}Q_{k+1}^{1} \mid P_k^{1}Q_k^{2}\right) & P\left(P_{k+1}^{1}Q_{k+1}^{2} \mid P_k^{1}Q_k^{2}\right) & P\left(P_{k+1}^{2}Q_{k+1}^{1} \mid P_k^{1}Q_k^{2}\right) & P\left(P_{k+1}^{2}Q_{k+1}^{2} \mid P_k^{1}Q_k^{2}\right) \\
P\left(P_{k+1}^{1}Q_{k+1}^{1} \mid P_k^{2}Q_k^{1}\right) & P\left(P_{k+1}^{1}Q_{k+1}^{2} \mid P_k^{2}Q_k^{1}\right) & P\left(P_{k+1}^{2}Q_{k+1}^{1} \mid P_k^{2}Q_k^{1}\right) & P\left(P_{k+1}^{2}Q_{k+1}^{2} \mid P_k^{2}Q_k^{1}\right) \\
P\left(P_{k+1}^{1}Q_{k+1}^{1} \mid P_k^{2}Q_k^{2}\right) & P\left(P_{k+1}^{1}Q_{k+1}^{2} \mid P_k^{2}Q_k^{2}\right) & P\left(P_{k+1}^{2}Q_{k+1}^{1} \mid P_k^{2}Q_k^{2}\right) & P\left(P_{k+1}^{2}Q_{k+1}^{2} \mid P_k^{2}Q_k^{2}\right)
\end{bmatrix} \\
&= \begin{bmatrix}
(1-r)^2 & r(1-r) & r(1-r) & r^2 \\
r(1-r) & (1-r)^2 & r^2 & r(1-r) \\
r(1-r) & r^2 & (1-r)^2 & r(1-r) \\
r^2 & r(1-r) & r(1-r) & (1-r)^2
\end{bmatrix}
\end{aligned} \tag{1}$$

De maneira similar, as matrizes de transição para $w=2,3,4$ podem ser expressas como

$$\mathbf{H}_{k,r}^{2} = \begin{bmatrix}
r(1-r) & (1-r)^2 & r^2 & r(1-r) \\
(1-r)^2 & r(1-r) & r(1-r) & r^2 \\
r^2 & r(1-r) & r(1-r) & (1-r)^2 \\
r(1-r) & r^2 & (1-r)^2 & r(1-r)
\end{bmatrix}$$

$$\mathbf{H}_{k,r}^{3} = \begin{bmatrix}
r(1-r) & r^2 & (1-r)^2 & r(1-r) \\
r^2 & r(1-r) & r(1-r) & (1-r)^2 \\
(1-r)^2 & r(1-r) & r(1-r) & r^2 \\
r(1-r) & (1-r)^2 & r^2 & r(1-r)
\end{bmatrix}$$

$$\mathbf{H}_{k,r}^{4} = \begin{bmatrix} r^2 & r(1-r) & r(1-r) & (1-r)^2 \\ r(1-r) & r^2 & (1-r)^2 & r(1-r) \\ r(1-r) & (1-r)^2 & r^2 & r(1-r) \\ (1-r)^2 & r(1-r) & r(1-r) & r^2 \end{bmatrix}.$$

Uma vez que a única maneira de acessar os genótipos é usando a informação dos fenótipos, é necessário definir matrizes indicadoras $\mathbf{I}_{z_{j,k}}$, que associam os valores fenotípicos $z_{j,k}$ observados dos marcadores M_k aos possíveis genótipos $x_{j,k}$ para o indivíduo j, selecionando apenas as transições possíveis. Estas matrizes $\mathbf{I}_{z_{j,k}}$ multiplicarão $\mathbf{H}_{k,r}^{w}$. Por exemplo, se o marcador é completamente informativo (tipo A), não há estados ocultos e as matrizes indicadoras serão

$$\mathbf{I}_{\{P_k^1 Q_k^1\}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \mathbf{I}_{\{P_k^1 Q_k^2\}} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{I}_{\{P_k^2 Q_k^1\}} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \mathbf{I}_{\{P_k^2 Q_k^2\}} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Entretanto, quando um fenótipo particular corresponde a dois ou mais possíveis genótipos, outras $\mathbf{I}_{z_{j,k}}$ são necessárias e podem ser obtidas simplesmente somando-se as matrizes correspondentes aos fenótipos envolvidos. Por exemplo, se $z_{j,k} = \{P_k^1 Q_k^1, P_k^1 Q_k^2\}$, $\mathbf{I}_{\{P_k^1 Q_k^1, P_k^1 Q_k^2\}} = \mathbf{I}_{\{P_k^1 Q_k^1\}} + \mathbf{I}_{\{P_k^1 Q_k^2\}}$. Também é necessário definir a distribuição do estado inicial, que é a distribuição de probabilidade incondicional ou *a priori* de x_k na população. Esta distribuição pode ser indicada por $\mathbf{q}' = \left[\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right]$. Portanto, dada uma sequência de fases de ligação entre os marcadores nos cromossomos dos genitores é possível calcular a verossimilhança das observações para uma determinada ordem de marcadores $M_1, \cdots, M_m$ usando

$$
\begin{aligned}
L &= \prod_{j=1}^{n} P\left(z_{j,1} \cdots z_{j,m} \mid w_1 \cdots w_{m-1}\right) \\
&= \prod_{j=1}^{n} \mathbf{q}' \, \mathbf{I}_{z_{j,1}} \mathbf{H}_{1,r}^{w} \, \mathbf{I}_{z_{j,2}} \mathbf{H}_{2,r}^{w} \cdots \mathbf{I}_{z_{j,m-1}} \mathbf{H}_{m-1,r}^{w} \, \mathbf{I}_{z_{j,m}} \mathbf{c}
\end{aligned} \tag{2}
$$

sendo n o tamanho da população experimental e $\mathbf{c} = \{1\}_{4\times 1}$.

O cálculo da verossimilhança na Equação 2 requer o valor de r nas matrizes $\mathbf{H}_{k,r}^{w}$. Este valor deve ser estimado baseado no algoritmo EM (BAUM et al., 1970; DEMPSTER; LAID; RUBIN, 1977). Para tanto, seja $M_i, \cdots, M_k, \cdots, M_l$ a sequência de marcadores no mapa, em que M_i e M_l são os marcadores completamente observáveis (tipo A) mais adjacentes à direita e a esquerda de M_k. Se não houver marcadores do tipo A, $M_i = M_1$ e $M_l = M_m$ podem ser usados (JIANG; ZENG, 1997). Para cada indivíduo j é necessário calcular $P\left(x_k x_{k+1} \mid z_i \cdots z_k \cdots z_l, w_i \cdots w_l\right)$, que é a probabilidade condicional dos genótipos dos dois locos adjacentes dada uma sequência correspondente de fenótipos e fases de ligação. Os valores para todas as combinações genotípicas de x_k e x_{k+1} podem ser representados em uma matriz $\mathbf{A}_k^w$ (4×4):

$$
\mathbf{A}_k^w = \frac{\left[\left(\mathbf{q} \circ \mathbf{p}_k^L\right)\left(\mathbf{p}_{k+1}^R\right)'\right] \circ \mathbf{H}_{k,r}^{w}}{\left(\mathbf{q} \circ \mathbf{p}_k^L\right)' \mathbf{H}_{k,r}^{w} \, \mathbf{p}_{k+1}^R} \tag{3}
$$

em que $\circ$ denota o produto de Hadamard e $'$ denota transposição de vetores ou matrizes. $\mathbf{p}_{k+1}^R$ é um vetor que contém as probabilidades de observar-se uma sequência de marcadores com fenótipos $z_{j,k+2}, \cdots, z_l$ à direita do marcador com genótipo x_{k+1} dada uma sequência de fases de ligação e pode ser calculado por

$$
\mathbf{p}_{k+1}^R = \mathbf{I}_{z_{j,k+1}} \mathbf{H}_{k+1,r}^{w} \, \mathbf{I}_{z_{j,k+2}} \mathbf{H}_{k+2,r}^{w} \cdots \mathbf{I}_{z_{j,l-1}} \mathbf{H}_{l-1,r}^{w} \, \mathbf{I}_{z_{j,l}} \mathbf{c}
$$

De maneira similar,

$$
\mathbf{p}_k^L = \mathbf{I}_{z_{j,k}} \mathbf{H}_{k-1,r}^{w} \, \mathbf{I}_{z_{j,k-1}} \mathbf{H}_{k-2,r}^{w} \cdots \mathbf{I}_{z_{j,i-1}} \mathbf{H}_{i,r}^{w} \, \mathbf{I}_{z_{j,i}} \mathbf{c}
$$

A fração de recombinação para o k-ésimo intervalo pode ser atualizada usando

$$r = \frac{1}{2n} \sum_{j=1}^{n} \mathbf{c}' \left(\mathbf{D}^w \circ \mathbf{A}_k^w \right) \mathbf{c} \tag{4}$$

em que n é o tamanho da população experimental e $\mathbf{D}^w$ é uma matriz 4×4 que contém o número esperado de *crossing overs* para todas as combinações de genótipos x_k e x_{k+1}:

$$\mathbf{D}^1 = \begin{bmatrix} 0 & 1 & 1 & 2 \\ 1 & 0 & 2 & 1 \\ 1 & 2 & 0 & 1 \\ 2 & 1 & 1 & 0 \end{bmatrix} \qquad \mathbf{D}^2 = \begin{bmatrix} 1 & 0 & 2 & 1 \\ 0 & 1 & 1 & 2 \\ 2 & 1 & 1 & 0 \\ 1 & 2 & 0 & 1 \end{bmatrix}$$

$$\mathbf{D}^3 = \begin{bmatrix} 1 & 2 & 0 & 1 \\ 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 2 \\ 1 & 0 & 2 & 1 \end{bmatrix} \qquad \mathbf{D}^4 = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 1 & 2 & 0 & 1 \\ 1 & 0 & 2 & 1 \\ 0 & 1 & 1 & 2 \end{bmatrix}.$$

O valor de r para cada intervalo pode ser estimado usando-se as Equações 3 e 4 até a convergência, quando o mapa e a verossimilhança multiponto são obtidos.

2.1.4.2 Fases de Ligação

Wu et al. (2002a) mostraram como obter as probabilidades das possíveis fases de ligação dos marcadores nos cromossomos dos genitores usando o teorema de Bayes na abordagem das análises de dois e três pontos. Assumindo uma distribuição *a priori* uniforme para as fases de ligação, os autores apresentaram o cálculo da probabilidade *a posteriori* das fases de ligação. Dessa forma utiliza-se a verossimilhança multiponto obtida na Equação 2 e a distribuição *a priori* uniforme para calcular a probabilidade de uma sequência de fases de ligação. Para cada intervalo, existem $4^{(m-1)}$ possíveis sequências de fases de ligação que devem ser avaliadas. Seja $\mathbf{w}_t$ um vetor de tamanho $m-1$ que contém uma das 4^{m-1} sequências de fases de ligação possíveis ($t = 1, \cdots, 4^{m-1}$). Para uma determinada ordem, a probabilidade de cada $\mathbf{w}_t$ nos cromossomos dos genitores dado seu fenótipo pode ser calculada por

$$P\left(\mathbf{w}_t \mid z_1 \cdots z_m\right) = \frac{\prod_{j=1}^{n} P\left(z_{j,1} \cdots z_{j,m} \mid \mathbf{w}_t\right)}{\sum_{t=1}^{4^{m-1}} \prod_{j=1}^{n} P\left(z_{j,1} \cdots z_{j,m} \mid \mathbf{w}_t\right)}$$

que é proporcional à verossimilhança que permite a comparação das $\mathbf{w}_t$ sequências.

Deve-se mencionar que este procedimento requer cálculo intensivo e só é viável para pequenos valores de m. A seguir, apresentaremos uma estratégia de mapeamento que utiliza a informação de dois pontos para reduzir o espaço de busca.

2.1.4.3 Estratégia de Mapeamento

A primeira tarefa na construção de um mapa genético é a atribuição de marcadores aos grupos de ligação. Esta tarefa é simples e pode ser feita utilizando a análise de dois pontos (LANDER et al., 1987) baseando-se num limiar para as estimativas de fração de recombinação e seus respectivos LOD Scores (*logarithm of odds ratio* - MORTON, 1955). O próximo passo é ordenar os marcadores dentro dos grupos de ligação formados, o que é considerado um caso especial do clássico problema do caixeiro viajante (DOERGE, 1996; MESTER et al., 2003; TAN; FU, 2006), que consiste na escolha da melhor ordem entre as $m!/2$ possíveis. No entanto, é impossível verificar todas as possíveis ordens quando m é grande. Dessa forma, vários algoritmos foram propostos para contornar este problema, tais como o algoritmo implementado no comando TRY do programa MAPMAKER/EXP (LANDER et al., 1987).

O algoritmo consiste na busca exaustiva pela ordem com maior verossimilhança entre $m^*!/2$ ordens ($m^* < m$, tal que $m^*!/2$ é um número viável de ordens a serem avaliadas). Para cada uma das $m^*!/2$ ordens obtidas, há 4^{m^*-1} possíveis sequências de fases de ligação que teriam de ser avaliadas. Obviamente, esta estratégia é adequada apenas para pequenos valores de m^*. Para reduzir o espaço de busca da sequência de fases de ligação mais provável, devem ser utilizadas as informações do teste de dois pontos, obtidas com base no método de Wu et al. (2002a). As probabilidades *a posteriori* das fases de ligação são calculadas para cada intervalo e apenas aquelas que ultrapassarem um determinado limiar serão consideradas para a análise multiponto. Por meio deste procedimento, o espaço de busca de fases de ligação pode ser muito reduzido e uma sequência de m^* marcadores ordenados pode ser obtida. Na etapa seguinte, um novo marcador é posicionado no início, no fim e entre todos os m^* marcadores ordenados. Em seguida

é verificada qual é a combinação de fase e posição que fornece a verossimilhança mais elevada, também considerando a informação de dois pontos. Este procedimento é repetido até que todos os marcadores no grupo de ligação sejam posicionados.

2.2 Mapeamento Genético em Autopoliploides

Espécies poliploides dividem-se basicamente em dois grupos: i) *alopoliploides* e ii) *autopoliploides*. Alopoliploides apresentam dois ou mais genomas derivados de espécies distintas enquanto que os autopoliploides apresentam múltiplos genomas originários da mesma espécie (SINGH, 2003). Uma vez que os alopoliploides envolvem a junção de dois ou mais genomas diferentes, espera-se que o pareamento dos cromossomos homólogos durante a meiose seja similar àquele observado em células diploides (SINGH, 2003). Logo, as metodologias usadas para a construção de mapas genéticos nessas espécies são as mesmas usadas para a construção de mapas em espécies diploides. Alguns exemplos dessas espécies são trigo, amendoim, café, aveia, algodão e festuca (GALLAIS, 2003). Entretanto, em autopoliploides existem mais de dois cromossomos homólogos que podem se parear durante a meiose. Esses cromossomos podem se dispor dois a dois formando *bivalentes* ou podem se dispor em configurações chamadas *multivalentes*. Esse comportamento origina um padrão complexo de herança chamado de *padrão polissômico*, o qual torna a construção de mapas genéticos nessas espécies bastante complexa (GALLAIS, 2003). Associados à herança polissômica, outros fatores dificultam a construção de mapas genéticos em autopoliploides como o pareamento preferencial e a dupla redução, um fenômeno no qual as cromátides irmãs migram para o mesmo gameta causando distorções de segregação sistemáticas (EDME; GLYNN; COMSTOCK, 2006; LUO et al., 2006)

A herança polissômica caracteriza-se pela complexidade de segregação observada em uma progênie. Por exemplo, considerando a autofecundação de um indivíduo autotetraploide com genótipo $AAaa$, a segregação esperada na progênie é $1/36$ AAAA : $2/9$ AAAa : $1/2$ AAaa : $2/9$ Aaaa : $1/36$ aaaa, enquanto que em uma espécie diploide com genótipo Aa essa proporção é $1/4$ AA : $1/2$ Aa : $1/4$ aa. Essa diferença ocorre porque em autotetraploides os cromossomos podem se parear de várias maneiras, enquanto que num diploide, existe apenas uma maneira possível (HALDANE, 1930). Devido a esse padrão de herança polissômica, os métodos para a construção de mapas genéticos em poliploides não apresentam os mesmos avanços observados em espécies diploides (LUO et al., 2006; LEACH et al., 2010). Os primeiros estudos sobre mapeamento e ligação em autopoliploides remontam às décadas de 30 e 40 (HALDANE, 1930;

MATHER, 1936; FISCHER, 1947). Entretanto, apenas com o surgimento da tecnologia de marcadores moleculares e com o avanço de ferramentas computacionais adequadas tornou-se possível a realização de estudos que levassem em conta a herança polissômica dessas espécies.

Wu et al. (1992) propôs um método para a construção de mapas genéticos em autopoliploides utilizando marcadores em dose única, chamados de *"single dose restriction fragments (SDRFs)"*. Neste trabalho assumiu-se que os cromossomos pareavam-se de maneira aleatória e que não foi considerado o fenômeno da dupla redução. Dessa forma, as frequências esperadas numa população originária do cruzamento de dois indivíduos autopoliploides e com o mesmo nível de ploidia, sendo que apenas um deles apresentam a marca em dose simples, será sempre 1:1 (presença : ausência), independentemente do nível de ploidia considerado. O método proposto por Wu et al. (1992) tem a grande vantagem de não depender do nível de ploidia para a obtenção das estimativas de fração de recombinação. Entretanto, Wu et al. (1992) mostraram que, para autopoliploides, essas estimativas apresentam erros padrão altos quando os marcadores encontram-se em repulsão. Dessa forma, verificou-se que um número muito grande de indivíduos era necessário para detectar ligação entre marcadores com este padrão de segregação. Além disso, em espécies autopoliploides, é necessário reunir os cromossomos mapeados em grupos de homologia, ou seja, agrupar os cromossomos que foram mapeados com seus homólogos de modo a totalizar o nível de ploidia da espécie.

2.2.1 Incorporação de Marcadores com Outras Doses nos Mapas Genéticos

Além dos marcadores em dose única, Da Silva (1993) propôs a utilização de marcadores em doses duplas e triplas, chamados de *"double dose restriction fragments (DDRFs)"* e *"triple dose restriction fragments (TDRFs)"*, respectivamente. Ao contrário dos marcadores em dose única, marcadores em doses duplas e triplas possibilitam acessar informações de um determinado loco em mais de um homólogo, ou seja, torna-se possível o estudo da dosagem alélica, a qual é extremamente importante em espécies poliploides (Figura 2).

Entretanto, a construção de mapas genéticos usando tais marcadores não é trivial, uma vez que os padrões de segregação obtidos são bastante complexos. Soma-se a isso o fato dos marcadores usualmente empregados (AFLPs, microssatélites) terem comportamento dominante nesse tipo de situação. O comportamento dominante de, por exemplo, um microssatélite em uma espécie autopoliploide ocorre devido à mistura de classes genotípicas que esses marcadores apresentam. Seja, por exemplo, um indivíduo auto-octaploide com dois alelos A e seis alelos a para

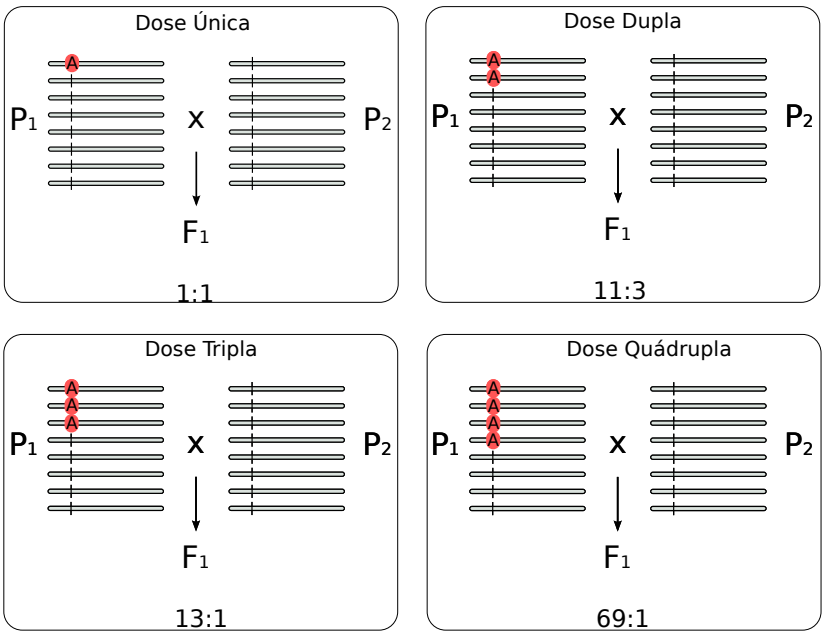

Figura 2 – Configuração dos marcadores em dose única, dupla, tripla e quádrupla nos grupos de homologia auto-octaploides de um dos genitores de um cruzamento biparental, resultando nos padrões de segregações 1:1, 11:3, 13:1 e 69:1

o mesmo loco (marcador em dose dupla). Se esse indivíduo for cruzado com um outro que contém apenas alelos a, podem existir três classes genotípicas na progênie: $AAaaaaaa$, $Aaaaaaaa$ e $aaaaaaaa$ nas proporções de $3 : 8 : 3$ (a obtenção dessas proporções será abordada no item Material e Métodos). Embora existam três classes genotípicas, a técnica molecular não permite separar as classes $AAaaaaaa$ e $Aaaaaaaa$, resultando em apenas duas classes na proporção de $11 : 3$ (presença e ausência).

O método desenvolvido por Da Silva (1993) e explicado em detalhes por Ripol et al. (1999) permite a utilização desse tipo de marcador e consiste no cálculo da probabilidade das classes genotípicas que surgem quando são considerados dois locos. Dada a natureza dominante dos marcadores utilizados, existem quatro classes genotípicas possíveis para dois locos, digamos, A e B: AB, A, B, e 0 (ausência dos dois alelos). Como as frequências dessas classes ocorrem em função da ligação entre os dois locos, podemos escrever as probabilidades de ocorrência de cada uma delas, $P(AB)$, $P(A)$, $P(B)$ e $P(0)$, em função da fração de recombinação e do nível de ploidia. Para a obtenção desses valores, são feitas diversas pressuposições como a ocorrência de pareamento aleatório e formação de bivalentes. Ripol et al. (1999) apresentam estas probabilidades para diversas dosagens e fases de ligação. Uma vez obtidas estas probabilidades em função da fração de recombinação, é possível montar a função de verossimilhança que é dada

por

$$L = P(AB)^{n_{AB}} P(A)^{n_A} P(B)^{n_B} P(0)^{n_0}$$

em que L é a verossimilhança do modelo, n_{AB}, n_A, n_B e n_0 são as contagens das classes genotípicas AB, A, B, e 0, respectivamente. Através dessa equação, é possível obter estimativas de máxima verossimilhança. Isto pode ser feito resolvendo-se a primeira derivada da função usando o algoritmo de Newton-Raphson, implementado por exemplo na função `optim` do software R (R DEVELOPMENT CORE TEAM, 2011).

Com o mapa genético multiponto construído usando o algoritmo descrito em 2.1.4, pode-se estimar a fração de recombinação entre os marcadores em dose única já posicionados no mapa e os marcadores de outras doses, ainda não posicionados. Para tanto faz-se necessário o cálculo das frações de recombinação entre marcadores em dose única e marcadores em outras doses. Como exemplo, considere dois marcadores A e B com fração de recombinação r. A Tabela 2 mostra o logaritmo das funções de verossimilhança para estes dois marcadores considerando diferentes padrões de segregação.

Estimadores de máxima verossimilhança podem ser obtidos para cada uma das funções da Tabela 2. Dessa forma as frações de recombinação podem ser calculadas entre todos os pares de marcadores. Para o teste de ligação entre os pares de marcadores, pode-se utilizar a equação a seguir considerando-se algum limiar de LOD para declarar ligação.

$$LOD = \log_{10} \frac{L(\hat{r} \mid \mathbf{n})}{L(r = 0.5 \mid \mathbf{n})}$$

sendo $\hat{r}$ a estimativa de máxima verossimilhança da fração de recombinação para dois marcadores em questão. Finalmente, os marcadores com doses superiores ligados aos marcadores em dose única podem ser posicionados no mapa multiponto previamente construído.

Devido à natureza dominante dos marcadores utilizados por Da Silva (1993) e por diversos outros trabalhos que aplicaram esse método (DA SILVA; SORRELLS, 1996; RIPOL et al., 1999; AITKEN; JACKSON; MCINTYRE, 2005; ANDRU et al., 2011), o poder estatístico para a detecção de ligação entre locos em múltiplas doses é muito baixa, especialmente quando os marcadores encontram-se em fase de repulsão, sendo necessário o uso de populações experimentais com tamanhos inviáveis. Devido a esta limitação, a grande maioria dos mapas de espécies autopoliploides, principalmente aquelas com ploidia alta, como é o caso da cana-de-açúcar, é construída usando marcadores em dose única (PASTINA et al., 2010).

Tabela 2 – Logaritmo das funções de verossimilhança para duplas de marcadores com diferentes padrões de segregação (m : nível de ploidia, r : fração de recombinação) (RIPOL et al., 1999)

Combinação de marcadores	Logaritmo da equação de verossimilhança $l(r)$
Simplex e Duplex (associação)	$n_{AB} \times \log\left(\frac{1}{2} - \frac{(m-2)r}{4(m-1)}\right) + n_A \times \log\left(\frac{(m-2)r}{4(m-1)}\right) +$ $n_B \times \log\left(\frac{(m-2)r}{4(m-1)} + \frac{m}{4(m-1)}\right) + n \times \log\left(\frac{m-2}{4(m-1)} - \frac{(m-2)r}{4(m-1)}\right)$
Simplex e Duplex (repulsão)	$n_{AB} \times \log\left(\frac{3m-4}{8(m-1)} - \frac{r}{2(m-1)}\right) + n_B \times \log\left(\frac{r}{2(m-1)} + \frac{3m}{8(m-1)}\right) +$ $n_A \times \log\left(\frac{m}{8(m-1)} - \frac{r}{2(m-1)}\right) + n \times \log\left(\frac{r}{2(m-1)} + \frac{m-4}{8(m-1)}\right)$
Simplex e Triplex (associação)	$n_{AB} \times \log\left(\frac{1}{2} - \frac{(m-4)r}{8(m-1)}\right) + n_A \times \log\left(\frac{(m-4)r}{8(m-1)}\right) +$ $n_B \times \log\left(\frac{(m-4)r}{8(m-1)} + \frac{3m}{8(m-1)}\right) + n \times \log\left(\frac{m-4}{8(m-1)} - \frac{(m-4)r}{8(m-1)}\right)$
Duplex e Duplex (associação)	$n_{AB} \times \log\left(\frac{(m-2)r^2}{4(m-1)} - \frac{(m-2)r}{2(m-1)} + \frac{3m-2}{4(m-1)}\right) +$ $(n_B + n_A) \times \log\left(\frac{(m-2)r^2}{4(m-1)} + \frac{(m-2)r}{2(m-1)}\right) +$ $n \times \log\left(\frac{(m-2)r^2}{4(m-1)} - \frac{(m-2)r}{2(m-1)} + \frac{m-2}{4(m-1)}\right)$

2.2.2 Mapeamento Genético em Autotetraploides

Os métodos genético-estatísticos usados para a construção de mapas em autopoliploides dividem-se claramente naqueles desenvolvidos para autotetraploides e em métodos desenvolvidos para espécies com maiores níveis de ploidia. A literatura científica relacionada aos métodos de autotetraploides vem apresentando grandes avanços na última década. Esses métodos valem-se da grande quantidade de informações sobre a teoria citogenética dessas espécies. Fenômenos como a dupla redução e pareamentos preferenciais são bastante estabelecidos em autotetraploides. Este fato, juntamente com o avanço da tecnologia computacional, torna tais conhecimentos prontamente disponíveis para o desenvolvimento de modelos genético-estatísticos adequados. Na realidade, os primeiros estudos realizados em autopoliploides por Haldane (1930), Mather (1936) e Fischer (1947) foram feitos considerando-se apenas espécies autotetraploides. Sved

(1962) relatou as relações entre frequências de recombinação entre diploides e autotetraploides. Nesse trabalho foi demonstrado que, considerando-se a formação de tetravalentes, o limite superior do espaço paramétrico da fração de recombinação é $3/4$, ao contrário de $1/2$ como observado em diploides. Os seja, diferentemente de diploides, o valor de fração de recombinação $1/2$ não indica necessariamente que os locos não estão ligados (SVED, 1962).

Uma das limitações para a construção de mapas genéticos em autopoliploides é a caracterização das classes genotípicas presentes em uma população de mapeamento (LUO; TAO; ZENG, 2000). Ao contrário dos padrões de segregação encontrados em diploides, espécies autopoliploides apresentam padrões complexos de segregação, muitas vezes difíceis de serem observados com as técnicas moleculares comumente utilizadas. As abordagens baseadas em marcadores em doses única, dupla e tripla proposta por Wu et al. (1992) e Da Silva (1993) não levam em conta codominância e multialelismo, logo, a detecção desses padrões complexos torna-se inviável. Luo, Tao e Zeng (2000) desenvolveram um modelo estatístico, aplicado a autotetraploides, para a predição dos genótipos em populações biparentais baseando-se no padrão dos fenótipos moleculares (padrão de bandas) observados na progênie, utilizando marcadores dominantes e codominantes. Nesse trabalho, os autores mostraram que foi possível alocar os indivíduos em classes genotípicas esperadas, sendo que esta classificação foi de fundamental importância para estudos posteriores incluindo análises de mapeamento e de QTLs (HACKETT; BRADSHAW; MCNICOL, 2001; LUO et al., 2001; MA et al., 2002; LUO; ZHANG; KEARSEY, 2004; LUO et al., 2006; LEACH et al., 2010).

Dentre os diversos estudos de mapeamento de QTLs existentes, destaca-se o estudo desenvolvido por Leach et al. (2010), o qual coloca o mapeamento genético de autotetraploides no mesmo nível das espécies diploides. Nesse estudo, os autores utilizaram a tecnologia de cadeias de Markov ocultas para a obtenção das estimativas de fração de recombinação e ordenação dos marcadores. Além disso, os autores consideraram fenômenos como a dupla redução e pareamentos preferenciais. A utilização das cadeias de Markov ocultas para a obtenção de estimativas contornam a propagação de erros ocasionados devido à falta de informação entre algumas combinações de marcadores (LANDER et al., 1987). No contexto de segregações complexas, essa técnica é particularmente importante, pois a cadeia permite que marcadores mais informativos forneçam informações para os demais, em função de sua proximidade (MOLLINARI et al., 2009).

Em linhas gerais, o método proposto por Leach et al. (2010) baseia-se na obtenção das probabilidades de transição entre genótipos de um genitor. No caso dos mapas genéticos, a

probabilidade de ocorrência de um determinado genótipo em um loco, digamos em um indivíduo tetraploide $BBbb$, está condicionado ao genótipo do loco adjacente, digamos $AAaa$. Dessa forma, dado que foi observado o genótipo $AAaa$, a probabilidade de ocorrência do genótipo $BBbb$ pode ser representada por $P(BBbb|AAaa)$. Uma vez que os dois locos são adjacentes, essa probabilidade é função da fração de recombinação (JIANG; ZENG, 1997). Dispondo-se essas probabilidades em uma matriz de transição tem-se (LEACH et al., 2010)

$$\begin{aligned}
\mathbf{T}(r) &= \begin{bmatrix}
P(B_2|A_2) & P(B_2|A_1) & P(B_2|A_1) & P(B_2|A_1) & P(B_2|A_1) & P(B_2|A_0) \\
P(B_1|A_2) & P(B_1|A_1) & P(B_1|A_1) & P(B_1|A_1) & P(B_1|A_1) & P(B_1|A_0) \\
P(B_1|A_2) & P(B_1|A_1) & P(B_1|A_1) & P(B_1|A_1) & P(B_1|A_1) & P(B_1|A_0) \\
P(B_1|A_2) & P(B_1|A_1) & P(B_1|A_1) & P(B_1|A_1) & P(B_1|A_1) & P(B_1|A_0) \\
P(B_1|A_2) & P(B_1|A_1) & P(B_1|A_1) & P(B_1|A_1) & P(B_1|A_1) & P(B_1|A_0) \\
P(B_0|A_2) & P(B_0|A_1) & P(B_0|A_1) & P(B_0|A_1) & P(B_0|A_1) & P(B_0|A_0)
\end{bmatrix} \\
&= \begin{bmatrix}
(1-r)^2 & \frac{(1-r)r}{2} & \frac{(1-r)r}{2} & \frac{(1-r)r}{2} & \frac{(1-r)r}{2} & r^2 \\
\frac{(1-r)r}{2} & (1-r)^2 & \frac{(1-r)r}{2} & \frac{(1-r)r}{2} & r^2 & \frac{(1-r)r}{2} \\
\frac{(1-r)r}{2} & \frac{(1-r)r}{2} & (1-r)^2 & r^2 & \frac{(1-r)r}{2} & \frac{(1-r)r}{2} \\
\frac{(1-r)r}{2} & \frac{(1-r)r}{2} & r^2 & (1-r)^2 & \frac{(1-r)r}{2} & \frac{(1-r)r}{2} \\
\frac{(1-r)r}{2} & r^2 & \frac{(1-r)r}{2} & \frac{(1-r)r}{2} & (1-r)^2 & \frac{(1-r)r}{2} \\
r^2 & \frac{(1-r)r}{2} & \frac{(1-r)r}{2} & \frac{(1-r)r}{2} & \frac{(1-r)r}{2} & (1-r)^2
\end{bmatrix}
\end{aligned}$$

em que $A_2 = AAaa$, $A_1 = Aaaa$, $A_0 = aaaa$, analogamente para B, e r é a fração de recombinação entre dois locos adjacentes. A matriz de transição para um genitor pode ser facilmente expandida para os dois genitores usando-se $\mathbf{T} \otimes \mathbf{T}$, em que $\otimes$ denota o produto de Kronecker de matrizes. Logo, a matriz terá dimensão 36×36. Os autores ainda mostram que esta matriz pode ser expandida para comportar dupla redução em um loco ou em ambos.

Baseado no modelo markoviano oculto, a verossimilhança para uma determinada ordem e um determinado vetor de frações de recombinação $\mathbf{r}$ pode ser obtida usando-se

$$\begin{aligned}
L(\mathbf{r}|\mathbf{X}) &= \prod_{i=1}^{n} Pr\left(\mathbf{X}_i|\mathbf{r}\right) \\
&= \prod_{i=1}^{n} Pr\left(x_{i1}, x_{i2}, \cdots, x_{il}|\mathbf{r}\right)
\end{aligned}$$

em que $\mathbf{X}$ denota a matriz de genótipos observados dos n indivíduos, $\mathbf{X}_i = (x_{i1}, x_{i2}, \cdots, x_{il})$ contém os genótipos observados do indivíduo i para todos os l locos e $\mathbf{r}$ é um vetor de frações de recombinação a ser estimado.

Expansões também foram feitas levando em conta o coeficiente de dupla redução adicionando-se um parâmetro relativo a esse coeficiente no modelo. Pode-se então estimar os parâmetros (no caso do modelo, o vetor $\mathbf{r}$) usando-se o algoritmo Baum-Welch (BAUM et al., 1970), que é um caso particular do algoritmo EM (DEMPSTER; LAID; RUBIN, 1977). Após o cálculo da verossimilhança, ela pode ser usada como critério para escolha do melhor modelo, ou seja, dentre os modelos avaliados escolhe-se aquele no qual os marcadores apresentam melhor ordem e as fases de ligação mais prováveis (MOLLINARI et al., 2009). Com isso completa-se a construção de um mapa genético em uma espécie autotetraploide. Ao contrário dos mapas obtidos com o método de Wu et al. (1992), que atualmente é o mais utilizado na literatura, ou com a integração utilizada por Garcia et al. (2006) e Oliveira et al. (2007), o mapa obtido usando-se este procedimento representa o grupo de homologia como um todo e não apenas uma porção dos cromossomos homólogos. É óbvio que isso é vantajoso por se tratar de espécies poliploides. Vale ressaltar que os modelos devem sempre procurar descrever o fenômeno em questão, ou seja, conceitualmente, os grupos de ligação devem ser assim modelados.

Entretanto, esse método foi desenvolvido para espécies autotetraploides e sua expansão para autopoliploides com maiores níveis de ploidia não é trivial por limitações estatísticas e computacionais. Por exemplo, no caso do cruzamento de dois indivíduos autododecaploides (nível de ploidia igual a 12), a dimensão da matriz $\mathbf{T} \otimes \mathbf{T}$, apresentada anteriormente, seria 853776×853776 totalizando aproximadamente 728 bilhões de valores, o que ocuparia o espaço na memória de um computador de aproximadamente 5,3 TB (terabytes). Esses valores claramente inviabilizam a expansão direta desse método. Outro problema enfrentado no mapeamento de autopoliploides com níveis de ploidia altos é o sistema de genotipagem adequado. Tais sistemas não são tão estabelecidos quanto àqueles utilizados em diploides e até mesmo em autotetraploides. Recentemente, Voorrips, Gort e Vosman (2011) desenvolveram um método que automatiza a atribuição dos fenótipos moleculares a classes genotípicas (“*genotype call*”) para sistemas de genotipagem que fornecem dois sinais de inensidade de um marcador bialélico. O método mostrou-se eficiente para a análise de dados em tetraploides, entretanto mostrou-se limitado quando utilizado em dados com altos níveis de ploidia (SERANG; MOLLINARI; GARCIA, 2012). Neste caso, as classes genotípicas são notoriamente difíceis de distinguir e muitas das vezes, as frequências esperadas em cada classe são diferentes entre si.

3 MATERIAL E MÉTODOS

3.1 Material

A população de mapeamento utilizada para exemplificar o uso do método aqui desenvolvido foi composta por 180 indivíduos originários do cruzamento entre as variedades pré-comerciais IACSP 95-3018 (genitor materno) e IACSP 93-3046 (genitor paterno). Os indivíduos da população foram genotipados com 241 locos SNP, 428 AFLPs (sendo 258 em dose única) e 674 microssatélites (sendo 376 em dose única). Os marcadores SNP foram genotipados usando a tecnologia Sequenom iPLEX MassARRAY® (OETH et al., 2007). Esta tecnologia é baseada na extensão de um iniciador alelo específico com um terminador de massa modificada (SEQUENOM, 2007). Os produtos dessa reação contendo o DNA foram analisados usando-se um espectrômetro de massa MALDI-TOF (*Matrix-Assisted Laser Desorption/Ionization-Time of Flight*) e cada região polimórfica de interesse foi detectada pela massa do alelo específico. Os dois genitores foram genotipados com 12 repetições para cada SNP. Se a eficiência de ionização for similar para os dois alelos, as intensidades produzidas pela espectrometria de massa são proporcionais a sua quantidade (com uma proporcionalidade muito constante se as reações foram realizadas na mesma amostra) (SERANG; MOLLINARI; GARCIA, 2012). Portanto, como a amplificação dos dois alelos é similar, espera-se que os desvios em relação às proporções esperadas sejam mínimos.

3.2 Métodos

3.2.1 Leitura e Classificação dos Dados de SNPs

Os dados obtidos com a plataforma de genotipagem Sequenom iPLEX MassARRAY® (OETH et al., 2007) fornecem dois sinais de intensidade de massa referentes ao loco SNP em questão. A intensidade de cada sinal é proporcional a dosagem de cada um dos alelos presentes no loco. Assim, se um loco auto-octaploide tiver o genótipo AACCCCCC (duas doses do nucleotídeo A e seis doses do nucleotídeo C), os sinais apresentarão proporção $\frac{2}{6}$. Esse comportamento torna os SNPs excelentes marcadores para estudos genéticos em poliploides. Quando esta técnica é utilizada na genotipagem de uma população biparental de uma espécie autopoliploide,

tona-se possível observar diversas classes genotípicas, diferentemente dos microssatélites. Por exemplo, se dois genitores auto-octaploides tiverem os genótipos AACCCCCC e AAACCCCC, espera-se que a progênie apresente seis classes genotípicas a saber: AAAAACCC, AAAACCCC, AAACCCCC, AACCCCCC, ACCCCCCC e CCCCCCCC. Para que todo potencial dessa tecnologia possa ser utilizado, é fundamental que métodos estatísticos adequados sejam desenvolvidos, uma vez que os dados observados devem ser associados a cada uma dessas classes (*"SNP genotype call"*). Essa não é uma tarefa trivial em autopoliplóides com níveis de ploidia altos. A seguir será apresentado um modelo Bayesiano desenvolvido no presente trabalho, o qual formaliza o *"SNP genotype call"* em autopoliploides incluindo a estimativa do nível de ploidia (SERANG; MOLLINARI; GARCIA, 2012).

3.2.1.1 Modelo Gráfico Bayesiano para Classificação dos SNPs

Foi utilizada a abordagem Bayesiana para a modelar a probabilidade dos dados observados dado um nível de ploidia e todos os genótipos simultaneamente. Com essa abordagem, foi possível modelar o processo gerador dos dados, permitindo então a construção do modelo a partir de suposições bastante realistas. A partir do modelo proposto foi realizada a inferência considerando-se todas as ploidias possíveis dentro de uma faixa pré-determinada e também todas as possibilidades de configuração dos genótipos. Assim foi possível escolher a configuração que maximiza a probabilidade a posteriori do modelo. Essa configuração é conhecida como configuração de *máxima posteriori ("Maximum a posteriori - MAP")* e garante que o resultado encontrado é o mais provável entre todos existentes naquela faixa de nível de ploidia. O método será apresentado sucintamente a seguir, dando especial atenção para a definição do modelo estatístico usado. Maiores detalhes do método, incluindo a inferência exata, podem ser obtidos em Serang, Mollinari e Garcia (2012).

Na Figura 3 é apresentado o modelo que considera uma população de mapeamento F_1 derivada do cruzamento de dois indivíduos heterozigóticos. As setas indicam dependências entre os nós, que são representados pelas circunferências. O conjunto de genótipos G depende da ploidia P. Essa dependência é dada em função de uma probabilidade condicional $\Pr(G|P)$. Partindo dessa ideia básica, outras conexões podem ser construídas.

Para o modelo gráfico apresentado na Figura 3, a *configuração genotípica* $G = (G_1, G_2, \ldots, G_n)$ é a coleção de classes genotípicas atribuídas a todos indivíduos do conjunto de dados. Seja o conjunto de possíveis genótipos para uma dada ploidia $\mu(P) =$

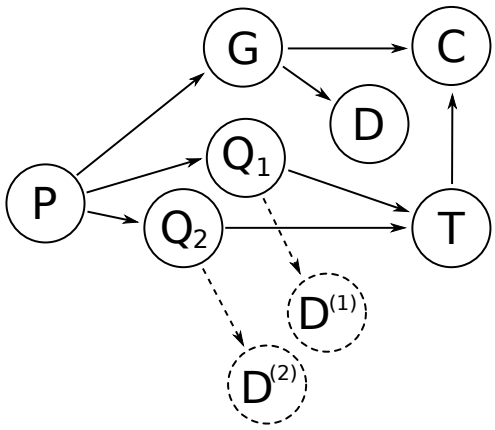

Figura 3 – Modelo gráfico Bayesiano que considera uma população de mapeamento F1

$\{\mu_0, \mu_1, \ldots \mu_P\}$. Por exemplo, para um loco diploide, $P = 2$ e o conjunto de possíveis genótipos é $\mu(P) = \{\mu_0 = (0, 2), \mu_1 = (1, 1), \mu_2 = (2, 0)\}$. Os dados observados em D são compostos por coleções de pontos $D_1, D_2, \ldots, D_n$, sendo que cada ponto D_i compreende um par de intensidades (x, y) para um indivíduo i. Assumindo que cada ponto depende apenas do indivíduo que o produziu a verossimilhança, para qualquer configuração de genótipos $G = g$, pode ser escrita como o produto de todos indivíduos

$$\Pr(D|G = g) = \prod_i \Pr(D_i|G_i = g_i)$$

Dessa forma, a verossimilhança pode ser modelada proporcionalmente a $\Pr(D_i|G_i = g_i)$ usando a distribuição normal com desvio padrão desconhecido σ

$$\Pr(D_i|G_i = g_i) \propto \frac{e^{-\frac{\|\widehat{D_i} - \widehat{g_i}\|_2^2}{\sigma^2}}}{\sigma^2}\sqrt{2\pi\sigma}$$

em que o operador $\hat{u} = \frac{u}{\|u\|_1}$ é usado para normalização de D_i e g_i com métrica L_1 (também conhecia como distância de Manhattan).

Para qualquer configuração genotípica $G = g$, o modelo calcula a distribuição dos possíveis genótipos $C = (C_0, C_1, \ldots C_P)$, sendo C_j igual a o número de indivíduos alocados na classe genotípica μ_j. A probabilidade de qualquer distribuição C pode ser modelada utilizando-se a distribuição teórica T dos genótipos. Dada a distribuição teórica da frequência dos genótipos na população $T = (p_1, p_2, \ldots p_P)$ sendo $p_1 + p_2 + \ldots p_P = 1$, a probabilidade de qualquer distribuição C ser observada é dada por

$$\Pr(C|T) = \frac{n!}{\prod_{j=0}^{P} C_j!} \prod_{j=0}^{P} p_j^{C_j}$$

A distribuição teórica dos genótipos na população biparental é modelada usando-se a distribuição hipergeométrica dos gametas. Denotemos $\mu_j.x$ a dosagem do primeiro alelo do par ordenado (x, y) e $\mu_j.y$ a dosagem do segundo alelo. Dados os genótipos dos genitores $Q_1 = q_1$ e $Q_2 = q_2$, ambos com valores em $\mu(p)$, a probabilidade de observar o gameta U_i originário de Q_1 é

$$\Pr(U_1 = u_1|Q_1 = q_1) = \frac{\binom{q_1.x}{u_1.x}\binom{q_1.y}{u_1.y}}{\binom{P}{u_1.x+u_1.y}}.$$

Portanto, a probabilidade de observar uma progênie μ_j é

$$p_j = \sum_{u_1,u_2:u_1.x+u_2.x=j} \Pr(U_1 = u_1|Q_1 = q_1) \Pr(U_2 = u_2|Q_2 = q_2) \tag{5}$$

Nesse modelo, os genótipos dos genitores Q_1 e Q_2, dependem da ploidia uma vez que o nível de ploidia de ambos deve estar contido em $\mu(P)$. A probabilidade *a priori* dos genótipos dos genitores, pode então ser modelada usando-se a distribuição uniforme pra o número de possíveis resultados $\Pr(Q_1, Q_2|P) = \binom{P+2}{2}$.

A Figura 3 apresenta setas e nós tracejados, os quais são considerados apenas se a informação dos genitores estiver disponível. A probabilidade desses parâmetros pode ser modelada através da independência condicional $\Pr(D|G)$:

$$\Pr(D^{(1)}|Q_1) = \prod_k \Pr(D_k^{(1)}|Q_1)$$

Quando os dados dos genitores são usados, o número de combinações únicas torna-se $(P+1) \times (P+1)$ e a distribuição uniforme é então usada considerando-se essas combinações.

Após a definição do modelo, deve ser feita a inferência de seus parâmetros. Um procedimento bastante usado para tanto é denominado *"grid search"*, que consiste em atribuir sequências de valores aos parâmetros verificando-se quais valores resultam no melhor modelo baseado em algum critério. No caso do presente estudo, este procedimento foi usado para estimar os parâmetros σ^2 e P e o critério levado em consideração foi a probabilidade *a posteriori*. Além da estimação do nível de ploidia e da variância, foi necessário verificar qual configuração genotípica (i.e. associação do dado observado a uma classe genotípica) resultou na maior probabili-

dade *a posteriori*. Dada a complexidade do problema, não foi possível aplicar métodos usuais de inferência usados em modelos gráficos uma vez que estes necessitariam o número de passos exponencial em n (n: número de indivíduos). Como $n \approx 200$, esse número é claramente inviável. Dessa forma, Serang, Mollinari e Garcia (2012) propuseram três métodos de inferência, sendo dois aproximados e um exato. Estes métodos não serão abordados neste texto, uma vez que esta é apenas uma etapa intermediária para a construção do mapa genético. Entretanto uma detalhada explicação dos modelos e procedimentos de estimação utilizados no presente trabalho podem ser obtidos em Serang, Mollinari e Garcia (2012). Para a realização dos procedimentos descritos acima, foi implementado um programa em linguagem Python chamado SuperMASSA que pode ser acessado no endereço `http:/statgen.esalq.usp.br/SuperMASSA/`. A classificação dos 241 SNPs foi feita utilizando-se o sistema operacional LINUX, kernel 3.2.0-6 em um computador com processador Intel® Core i7 com 2.8 GHz, 8MB de memória cache e 16 GB de memória RAM, usando os 8 núcleos disponíveis neste tipo de processador.

3.2.2 Classificação dos marcadores microssatélites e AFLPs

Embora diversas metodologias tenham sido propostas para a classificação de marcadores microssatélites e ALFPs quanto à dosagem, todas elas se baseiam no conhecimento prévio do nível de ploidia da espécie (RIPOL et al., 1999; DOERGE; CRAIG, 2000; BAKER; JACKSON; AITKEN, 2010). Como o exemplo utilizado no presente trabalho é uma população originária do cruzamento entre dois cultivares pré-comerciais de cana-de-açúcar, não é possível utilizar tais metodologias, uma vez que o nível de ploidia dos indivíduos é desconhecido. Ainda, como estes tipos de marcador não apresentam dois sinais de intensidade, como é o caso dos SNPs, não foi possível utilizar o método descrito no item anterior para estimação de doses e ploidia. Dessa forma, foram utilizados apenas marcadores em dose única, que foram selecionados usando-se o teste de qui-quadrado com nível de significância conjunto obtido através da correção de Bonferroni. Como assumiu-se o nível de significância individual de 0.05, o nível de significância conjunto, considerando-se as 674 marcas microssatélites e as 428 marcas ALFP, foi de $4,5 \times 10^{-5}$.

3.3 Desenvolvimento do modelo estatístico

O presente item apresenta em detalhes o novo método aqui desenvolvido. Dadas as exce-

lentes propriedades intrínsecas aos modelos markovianos ocultos (LANDER; GREEN, 1987; JIANG; ZENG, 1997; MOLLINARI et al., 2009), optou-se por modelar o fenômeno utilizando-se esta abordagem. Vale ressaltar que se forem considerados apenas dois marcadores em um modelo Markoviano oculto, o problema se reduz a uma análise de dois pontos, logo a implementação desse método permite a utilização de algoritmos tanto baseados em testes de dois pontos quanto algoritmos multiponto.

3.3.1 Notação

Considere uma população de mapeamento derivada do cruzamento entre dois indivíduos autopoliplóides P e Q com o mesmo nível de ploidia (família de irmãos completos). O nível de ploidia será denotado por m podendo ser qualquer número par maior do que zero. Sejam os vetores $\mathcal{P}_k^m = \{P_k^i\}$ e $\mathcal{P}_{k+1}^m = \{P_{k+1}^i\}$, e $\mathcal{Q}_k^m = \{Q_k^i\}$ e $\mathcal{Q}_{k+1}^m = \{Q_{k+1}^i\}$, $i = 1 \cdots m$ que denotam dois locos multi-alélicos adjacentes (k e $k+1$) em P e Q, respectivamente. O sobrescrito i indica um dos alelos do conjunto de possíveis alelos para os locos e cada loco possui m alelos em cada genitor. Por exemplo, no cruzamento entre dois indivíduos auto-hexaploides, $\mathcal{P}_k^6 = \{P_k^1, P_k^2, \cdots, P_k^6\}$, analogamente para $\mathcal{P}_{k+1}^6$, $\mathcal{Q}_k^6$ e $\mathcal{Q}_{k+1}^6$ (Figura 4). Os alelos denotados pelo mesmo número (sobrescrito) estão ligados (e.g. P_k^1 e P_{k+1}^1, e assim por diante). O número para o loco k também é usado para denotar o cromossomo homeólogo correspondente.

Genitor P | Genitor Q

$$\begin{array}{|c|c|c|c|c|c|} P_k^1 & P_k^2 & P_k^3 & P_k^4 & P_k^5 & P_k^6 \\ P_{k+1}^1 & P_{k+1}^2 & P_{k+1}^3 & P_{k+1}^4 & P_{k+1}^5 & P_{k+1}^6 \end{array} \times \begin{array}{|c|c|c|c|c|c|} Q_k^1 & Q_k^2 & Q_k^3 & Q_k^4 & Q_k^5 & Q_k^6 \\ Q_{k+1}^1 & Q_{k+1}^2 & Q_{k+1}^3 & Q_{k+1}^4 & Q_{k+1}^5 & Q_{k+1}^6 \end{array}$$

Figura 4 – Configuração de um cruzamento entre dois indivíduos auto-hexaploides ($m = 6$)

Considere ainda as seguintes pressuposições: i) formação exclusiva de bivalentes, ou seja, os cromossomos apenas se agrupam aos pares na meiose; ii) não há pareamento preferencial durante a formação de bivalentes; iii) todos os bivalentes apresentam a mesma fração de recombinação; iv) os bivalentes se comportam de maneira independente; v) há a separação das cromátides irmãs durante a meiose II. As suposições i e v garantem que não há dupla redução durante a meiose.

3.3.2 Formação dos Bivalentes

Durante a meiose I, mais especificamente na metáfase I, ocorre a formação dos bivalentes, isto é, a forma como os homólogos se posicionam. Em células diploides há apenas uma configuração de pareamento possível, ou seja, dois homólogos se pareiam para formar um bivalente. Entretanto, em células autopoliplóides, dadas as suposições anteriores, há mais de uma configuração possível. O número de bivalentes em cada configuração varia de acordo com o nível de ploidia. Considerando o nível de ploidia m, cada configuração possui $\frac{m}{2}$ bivalentes. O número de possíveis configurações de pareamentos, i.e., o número de possíveis pareamentos cromossômicos durante a meiose é

$$w_m = \frac{1}{\frac{m}{2}!} \prod_{i=2,4,\cdots,m} \binom{i}{2}$$

Analogamente à formação de um gameta em diploides para diferentes homólogos, a orientação dos bivalentes não tem efeito nas frequências esperadas de cada tipo de gameta, e portanto não precisa ser considerada. Por exemplo, para um grupo de homologia de um indivíduo autotetraploide (homólogos 1, 2, 3 e 4), há dois bivalentes e três configurações possíveis: homólogos se pareiam 1 com 2 e 3 com 4 ou 1 com 3 e 2 com 4 ou 1 com 4 e 2 com 3. $\Psi = \{\psi_j\}$, $j = 1, \cdots, w_m$ denota todas as possíveis configurações bivalentes para um determinado nível de ploidia m.

3.3.3 Frequências Gaméticas Esperadas para uma Dada Configuração de Bivalentes

Primeiramente, serão apresentadas as frequências gaméticas considerando-se apenas um genitor. Posteriormente, esses conceitos serão expandidos para dois genitores. Cada um dos bivalentes obtidos para uma dada configuração ψ_j pode resultar em dois tipos de cromossomos para os locos k e $k+1$: i) *não recombinantes*, que são resultado de bivalentes com zero ou qualquer número ímpar de recombinações entre os locos k e $k+1$; ii) *recombinantes*, que são resultado de bivalentes com qualquer número par e maior do que zero de recombinações. As probabilidades de todos os tipos cromossômicos originários de um único bivalente, podem ser

representadas na matriz

$$\mathbf{V} = \begin{bmatrix} \mathrm{P}(P_k^i, P_{k+1}^i|\psi_j) & \mathrm{P}(P_k^i, P_{k+1}^{i'}|\psi_j) \\ \mathrm{P}(P_k^{i'}, P_{k+1}^i|\psi_j) & \mathrm{P}(P_k^{i'}, P_{k+1}^{i'}|\psi_j) \end{bmatrix} = \begin{bmatrix} \frac{(1-r_k)}{2} & \frac{r_k}{2} \\ \frac{r_k}{2} & \frac{(1-r_k)}{2} \end{bmatrix}$$

em que r_k é a fração de recombinação entre os locos k e $k+1$, $i \neq i'$. Para uma dada configuração ψ_j, as frequências esperadas para todos os gametas (todos os tipos cromossômicos para todos os bivalentes) podem ser obtidas usando

$$\mathbf{H}_m^{\psi_j} = \mathbf{V}_1 \otimes \mathbf{V}_2 \otimes \cdots \otimes \mathbf{V}_{\frac{m}{2}}$$

em que $\otimes$ denota o produto de Kronecker de matrizes e os subscritos em $\mathbf{V}$ indicam os bivalentes correspondentes.

É fácil notar que $\mathbf{H}_m^{\psi_j}$ terá dimensões muito grandes para altos níveis de ploidia. Entretanto, simplificações são possíveis, já que todos os elementos de $\mathbf{H}_m^{\psi_j}$ são da forma

$$\Pr(\mathbf{p}_k^m,\ \mathbf{p}_{k+1}^m|\psi_j) = \frac{(1-r_k)^{\frac{m}{2}-l}(r_k)^l}{2^{\frac{m}{2}}} \tag{6}$$

em que os vetores $\mathbf{p}_k^m$ e $\mathbf{p}_{k+1}^m$ denotam um subconjunto de $\frac{m}{2}$ alelos presentes em $\mathcal{P}_k^m$ e $\mathcal{P}_{k+1}^m$, respectivamente; $\left\{\mathbf{p}_k^m, \mathbf{p}_{k+1}^m\right\}$ indica um gameta, l denota o número de cromossomos recombinantes entre os locos k e $k+1$, $l = 0, \cdots, \frac{m}{2}$.

Uma vez que foi assumido que os alelos com o mesmo número estão ligados, pode-se obter l pela simples verificação de $\left\{\mathbf{p}_k^m, \mathbf{p}_{k+1}^m\right\}$, e assim a Expressão 6 depende apenas de r_k, que é justamente o parâmetro a ser estimado. Por exemplo, para $\psi_1 = \{(1,2),(3,4),(5,6)\}$ ($m = 6$, Figura 5) um possível gameta seria formado por $\mathbf{p}_k^6 = \{P_k^1, P_k^3, P_k^5\}$ e $\mathbf{p}_{k+1}^6 = \left\{P_{k+1}^1, P_{k+1}^4, P_{k+1}^6\right\}$, tendo dois cromossomos recombinantes ($l = 2$). Obviamente, certos tipos de gametas não podem ser obtidos a partir de ψ_1, por exemplo, $P_k^1, P_k^2, P_k^5/P_{k+1}^1, P_{k+1}^2, P_{k+1}^5$, entretanto, eles estarão contidos em outros ψ_j.

P_k^1	P_{k+1}^1	P_k^3	P_{k+1}^3	P_k^5	P_{k+1}^5
P_k^2	P_{k+1}^2	P_k^4	P_{k+1}^4	P_k^6	P_{k+1}^6

Figura 5 – Uma possível configuração de pareamento em um auto-hexaploide, denominada ψ_1.

3.3.4 Frequências Gaméticas para Todas Configurações

Em situações reais, ψ_j não é conhecido, então os elementos de $\mathbf{H}_m^{\psi_j}$ precisam ser estimados para qualquer os possíveis ψ_j. É fundamental notar que $\Pr(\mathbf{p}_k^m,\ \mathbf{p}_{k+1}^m|\psi_j)$ tem a mesma forma para todos os ψ_j. Dessa forma, o problema se reduz a contar o número de possíveis ψ_j que podem resultar no genótipo observado do gameta. Todo gameta pode ter l variando de zero a $\frac{m}{2}$ cromossomos recombinantes. Se $l = 0$, não há informação sobre os outros cromossomos homólogos que migram para o outro polo da célula na anáfase I. Mas, uma vez que a ploidia é m, há $\frac{m}{2}!$ possibilidades de pareamento, então, há $\frac{m}{2}!$ possíveis ψ_j. Para $l > 0$, há $\frac{m}{2} - l$ cromossomos não recombinantes, então há $\left(\frac{m}{2} - l\right)!$ possibilidades de pareamento. Para os l recombinantes que restaram, o número de possíveis pareamentos é $l!$. Portanto, o número total de possíveis configurações de pareamento é $l!\left(\frac{m}{2} - l\right)!$. A Figura 6 apresenta alguns exemplos de todos os ψ_j possíveis para alguns gametas auto-octaploides.

A partir disso, segue-se que

$$\Pr(\mathbf{p}_k^m,\ \mathbf{p}_{k+1}^m) = \sum_{j=1}^{l!\left(\frac{m}{2}-l\right)!} \Pr(\mathbf{p}_k^m,\ \mathbf{p}_{k+1}^m|\psi_j)\Pr(\psi_j) \tag{7}$$

Dadas as pressuposições, a formula geral para qualquer l é

$$\Pr(\mathbf{p}_k^m,\ \mathbf{p}_{k+1}^m) = \frac{l!\left(\frac{m}{2} - l\right)!}{w_m}\frac{(1-r_k)^{\frac{m}{2}-l}(r_k)^l}{2^{\frac{m}{2}}}$$

3.3.5 Probabilidades de Transição

As probabilidades de transição $\mathrm{P}(\mathbf{p}_{k+1}^m|\mathbf{p}_k^m)$, ou seja, probabilidades condicionais dos genótipos dos gametas no loco $k+1$ dados o genótipo do gameta no loco k podem ser calculadas usando-se

$$\Pr(\mathbf{p}_{k+1}^m|\mathbf{p}_k^m) = \frac{\mathrm{P}(\mathbf{p}_k^m,\ \mathbf{p}_{k+1}^m)}{\mathrm{P}(\mathbf{p}_k^m)}$$

Após algumas simplificações algébricas apresentadas no Apêndice 1, tem-se

$$\Pr(\mathbf{p}_{k+1}^m|\mathbf{p}_k^m) = \frac{(1-r_k)^{\frac{m}{2}-l}(r_k)^l}{\binom{\frac{m}{2}}{l}} \tag{8}$$

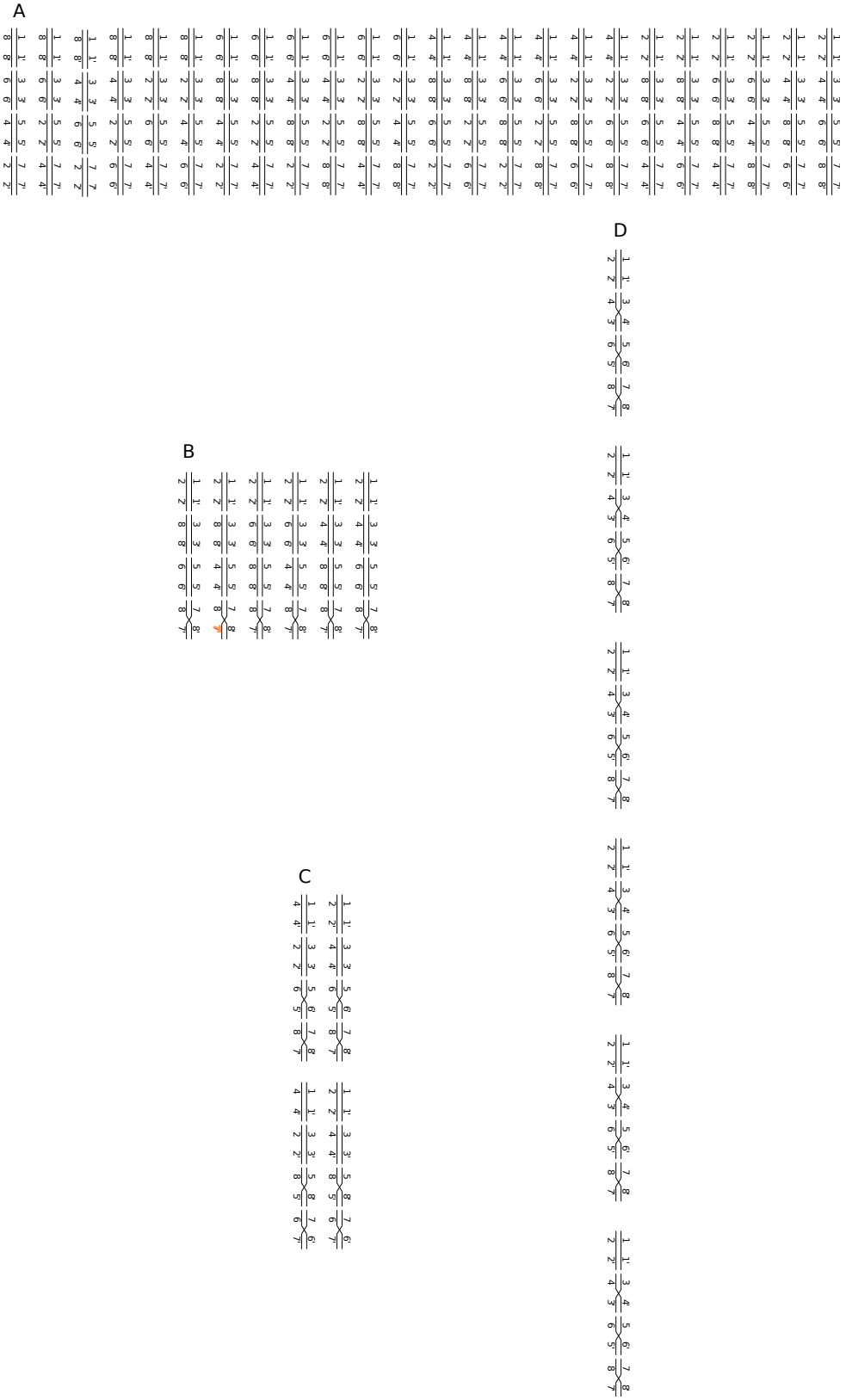

Figura 6 – Ilustração de todos os possíveis ψ_j para alguns gametas auto-octaploides ($m = 8$). A: $P_k^1, P_k^3, P_k^5, P_k^7/P_{k+1}^1, P_{k+1}^3, P_{k+1}^5, P_{k+1}^7$; B: $P_k^1, P_k^3, P_k^5, P_k^7/P_{k+1}^1, P_{k+1}^3, P_{k+1}^5, P_{k+1}^8$; C: $P_k^1, P_k^3, P_k^5, P_k^7/P_{k+1}^1, P_{k+1}^3, P_{k+1}^6, P_{k+1}^8$ e D: $P_k^1, P_k^3, P_k^5, P_k^7/P_{k+1}^1, P_{k+1}^4, P_{k+1}^6, P_{k+1}^8$. Nota-se que o número de possíveis ψ_j é $l!\left(\frac{m}{2}-l\right)!$ para todas as situações, dependendo apenas de l. A notação foi simplificada indicando apenas os alelos nos cromossomos.

Esta equação será usada na obtenção de estimativas de máxima verosimilhança de r_k, usando-se o modelo de Markov oculto. Esta equação reduz a necessidade de processamento computacional no cálculo da cadeia de Markov, uma vez que cada elemento da matriz de transição pode ser facilmente obtido, dependendo apenas de l. É interessante notar que fazendo-se $m = 4$ obtém-se a matriz de transição sem dupla redução apresentada por Hackett (2001) e depois expandida por Leach et al. (2010), mostrando que de fato foi apresentada aqui uma generalização para qualquer nível de ploidia. Obviamente a Equação 8 também serve para diploides, uma vez que fazendo-se $m = 2$, obtém-se a matriz de probilidades de transição (1), página 30.

3.3.6 Redução da Dimensão da Matriz de Transição

Seja $\mathbf{\Theta}_{P,k}^{m}$ um vetor contendo todos os $\binom{m}{\frac{m}{2}}$ possíveis $\mathbf{p}_k^m$. Por exemplo, para um autotetraploide, $\mathbf{\Theta}_{P,k}^{4} = \{(1,2),(1,3),(1,4),(2,3),(2,4),(3,4)\}$. Dependendo do sistema de marcador utilizado, algumas dessas classes genotípicas não podem ser distinguidas entre si. Por exemplo, para um marcador bialélico, como é o caso dos SNPs, uma possível codificação para um marcador em dose dupla seria $1 = A$, $2 = A$, $3 = a$, and $4 = a$. Portanto, $\mathbf{\Theta}_{P,k}^{4} = \{(A,A),(A,a),(A,a),(A,a),(A,a),(a,a)\}$. Enumerando-se as classes genotípicas no vetor $\tilde{\mathbf{\Theta}}_{P,k}^{m}$ (no exemplo, $\tilde{\mathbf{\Theta}}_{P,k}^{4} = \{2,1,1,1,1,0\}$), obtém-se a matriz de incidência $\mathbf{I}_k$ que é dada por

$$\mathbf{I}_{k_{i,j}} = \begin{cases} 1 & \text{se } j = \frac{m}{2} + 1 - \left(\tilde{\mathbf{\Theta}}_{P,k}^{m}\right)_i \\ 0 & \text{caso contrário} \end{cases}$$

em que o número de linhas de $\mathbf{I}_k$ é igual a $\binom{m}{\frac{m}{2}}$ e o número de colunas é igual ao número de elementos distintos em $\mathbf{\Theta}_{P,k}^{m}$. No exemplo anterior

$$\mathbf{I}_k = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Cada coluna de $\mathbf{I}_k$ representa um genótipo diferente (no exemplo, AA, Aa and aa) e cada

linha indica um dos $\binom{m}{\frac{m}{2}}$ possíveis gametas. Uma matriz $\mathbf{I}_{k+1}$ similar pode ser escrita para $\mathbf{p}^m_{k+1}$. Obviamente, o marcador presente no loco $k+1$ pode ter outro tipo de sistema de marcador, por exemplo, dominante. Sendo $\mathbf{T}(r_k)$ a matriz de transição obtida a partir da Equação 8, pode-se reduzir sua dimensão dependendo da informação contida no marcador usando

$$\tilde{\mathbf{T}}(r_k) = \mathbf{I}_k \left(\frac{\mathbf{1I'}_{k+1}}{\mathbf{I'}_k \mathbf{I}_k \mathbf{1I'}_{k+1}} \circ \mathbf{T}(r_k) \mathbf{I'}_{k+1} \right)$$

em que $\circ$ denota o produto direto de matrizes e $\mathbf{1}$ é uma matriz quadrada conforme na qual todos os elementos são iguais a um.

Supondo-se as mesmas frações de recombinação em ambos os genitores, pode-se incluir o genitor Q nas análises de forma análoga ao genitor P e escrevendo-se uma nova matriz de incidência $\mathbf{I}_k$ usando $\tilde{\mathbf{\Theta}}^m_{P,k} \oplus \tilde{\mathbf{\Theta}}^m_{Q,k}$ ($\oplus$ denota a soma de Kronecker de matrizes). Analogamente para $\mathbf{I}_{k+1}$. Então, a forma reduzida da matriz considerando-se os dois genitores é

$$\tilde{\mathbf{T}}(r_k) = \mathbf{I}_k \left\{ \frac{\mathbf{1I'}_{k+1}}{\mathbf{I'}_k \mathbf{I}_k \mathbf{1I'}_{k+1}} \circ \left(\mathbf{T}(r_k) \otimes \mathbf{T}(r_k) \right) \mathbf{I'}_{k+1} \right\} \tag{9}$$

A matriz a seguir mostra um exemplo da utilização da equação 9 na obtenção da matriz de transição considerando-se dois genitores autotetraploides. Tanto o marcador presente no loco k, quanto o marcador presente no loco $k+1$ são bialélicos e estão em dose dupla

$$\tilde{\mathbf{T}}(r_k) = \begin{bmatrix} r^4 & 4\,(1-r)\,r^3 & 6\,(1-r)^2\,r^2 & 4\,(1-r)^3\,r & (1-r)^4 \\ \frac{(1-r)^3\,r}{2} & 2\,(1-r)^2\,r^2 + (1-r)^3\,r + (1-r)^4 & \frac{5\,(1-r)\,r^3}{2} + 2\,(1-r)^2\,r^2 + \frac{5\,(1-r)^3\,r}{2} & r^4 + (1-r)\,r^3 + 2\,(1-r)^2\,r^2 & \frac{(1-r)\,r^3}{2} \\ (1-r)^2\,r^2 & 2\,(1-r)\,r^3 + 2\,(1-r)^3\,r & r^4 + 4\,(1-r)^2\,r^2 + (1-r)^4 & 2\,(1-r)\,r^3 + 2\,(1-r)^3\,r & (1-r)^2\,r^2 \\ \frac{(1-r)\,r^3}{2} & r^4 + (1-r)\,r^3 + 2\,(1-r)^2\,r^2 & \frac{5\,(1-r)\,r^3}{2} + 2\,(1-r)^2\,r^2 + \frac{5\,(1-r)^3\,r}{2} & 2\,(1-r)^2\,r^2 + (1-r)^3\,r + (1-r)^4 & \frac{(1-r)^3\,r}{2} \\ (1-r)^4 & 4\,(1-r)^3\,r & 6\,(1-r)^2\,r^2 & 4\,(1-r)\,r^3 & r^4 \end{bmatrix}$$

É interessante notar que a redução da dimensão da matriz de transição é análoga ao procedimento apresentado para diploides por Jiang e Zeng (1997) e Wu et al. (2002a). Entretanto as matrizes de incidência eram explícitas e apresentavam dimensões reduzidas (4×4, no máximo). Evidentemente, esta operação matricial torna-se inviável para níveis de ploidia altos uma vez que as dimensões das matrizes tornam-se muito grandes. Dessa forma foi desenvolvido um algoritmo computacional para a realização da redução das matrizes. Em linhas gerais, o algoritmo divide a matriz em subconjuntos e realiza contagens de elementos que devem ser somados. Como o número possível desses elementos é muito inferior às dimensões da matriz, esse cálculo torna-se viável.

Finalmente, a equação de verossimilhança pode ser obtida usando-se o procedimento apre-

sentado por Rabiner (1989) e aplicado no contexto de mapeamento genético por Jiang e Zeng (1997). Nesse procedimento são necessários três elementos básicos: i) função de *iniciação*, que é dada pela distribuição hipergeométrica, como exposto na página 46, Equação 5; ii) função de *emissão* e iii) *probabilidades de transição*. Esses dois últimos elementos já estão contidos no procedimento de redução de dimensão descrito anteriormente. Dessa forma, a equação de verossimilhança é dada por

$$L = \prod_{l=1}^{n} \mathbf{q}'\tilde{\mathbf{T}}(r_1)\tilde{\mathbf{T}}(r_2), \ldots, \tilde{\mathbf{T}}(r_{l-1})\mathbf{c} \tag{10}$$

sendo n o número de indivíduos, l o número de locos, $\mathbf{q}$ o vetor com as probabilidades de iniciação, $\mathbf{r} = \{r_1, r_2, \ldots, r_{l-1}\}$ o vetor de frações de recombinação, $\mathbf{c}$ um vetor com todos elementos iguais a um e $'$ denota transposição de matrizes. Esta verossimilhança pode ser usada como critério para a comparação de diferentes ordens (LANDER; GREEN, 1987). O valor de $\mathbf{r}$ pode ser estimado usando-se o algoritmo EM (DEMPSTER; LAID; RUBIN, 1977) com mostrado por Jiang e Zeng (1997), procedimento que foi aqui adotado.

As fases de ligação podem ser estimadas alterando-se os valores que cada classe assume em $\mathbf{\Theta}^m_{P,k}$ e $\mathbf{\Theta}^m_{P,k+1}$ (analogamente para o genitor Q) e comparando-se as verossimilhanças obtidas de maneira similar à apresentada na seção 2.1.4.2. Outra configuração de fases que poderia ser testada no exemplo anterior é $1 = A$, $2 = a$, $3 = A$, e $4 = a$. Portanto, $\mathbf{\Theta}^4_{P,k} = \{(A,a),(A,A),(A,a),(A,a),(a,a),(A,a)\}$ e $\tilde{\mathbf{\Theta}}^4_{P,k} = \{1,2,1,1,0,1\}$. Fazendo isso para todas as combinações possíveis, escolhe-se aquela que apresenta a maior verossimilhança. Embora o número de combinações possa ser muito grande quando o marcador utilizado é multialélico, para marcadores bialélicos, como é o caso dos SNPs, o número de combinações a serem testadas entre dois marcadores é $\min\{d_k, d_{k+1}\}+1$, sendo d_k e d_{k+1} as dosagens nos marcadores em questão. Por exemplo, se o nível ploidia dos locos for 16, existirão no máximo 17 possibilidades de configuração de fases de ligação, o que implica numa análise computacionalmente viável. O algoritmo de mapeamento descrito acima foi implementado no software R e as partes computacionalmente intensivas foram implementadas em linguagem C.

Para exemplificar o método proposto no presente trabalho, foram utilizados os marcadores SNP classificados como hexaploide e octaploide, além dos marcadores AFLPs e microssatélites com dose única. Primeiramente, testou-se a ligação entre todos os pares de marcadores SNPs classificados como hexaploides e marcadores AFLP e microssatélites em dose única. O mesmo procedimento foi realizado para os marcadores SNPs classificados como octaploides. O teste

de ligação foi baseado na estatística LOD Score com o uso da função de verossimilhança dada pela equação 10. Foi usado o limiar de 3 para declarar ligação. Após a obtenção dos grupos de ligação, os marcadores foram ordenados usando-se o algoritmo RCD (DOERGE, 1996) com inclusão das fases de ligação. Após a obtenção da ordem dos marcadores, o mapa foi reconstruído usando-se o método multiponto proposto neste trabalho.

4 RESULTADOS

4.1 Classificação dos Locos SNPs

Foram analisados 241 SNPs, sendo três deles selecionados para exemplificar o método de classificação (Figura 7). Os gráficos superiores mostram a distribuição dos dados brutos sem classificação. O eixo horizontal indica a intensidade do alelo (nucleotídeo) com menor massa e o eixo vertical indica a intensidade do alelo com maior massa. Cada gráfico contém 180 pontos que representam os indivíduos da população biparental. O primeiro gráfico apresenta claramente três nuvens de pontos. O segundo gráfico apresenta entre três e seis nuvens e o terceiro apresenta uma nuvem bastante dispersa. A classificação obtida com a utilização do software SUPERMASSA é apresentada na segunda linha de gráficos. Cada cor representa uma classe genotípica. Foram avaliados 8 níveis de ploidia (6, 8, 10, 12, 14, 16, 18 e 20), sendo escolhido o mais provável.

O primeiro loco foi classificado como decaploide contendo zero doses do nucleotídeo G no primeiro genitor e duas doses no segundo genitor. No caso de um loco decaploide, isso equivale a dez doses do nucleotídeo presente nos outros homólogos (no caso, T) do primeiro genitor e oito doses no segundo genitor. Ainda para este loco, os genótipos da progênie foram distribuídos em três classes, indicadas pelas cores vermelha, verde e azul. O segundo loco, foi classificado como dodecaploide com 4 doses do nucleotídeo A no primeiro genitor e duas doses no segundo genitor, com seis classes genotípicas na progênie. O terceiro loco, também dodecaploide, apresentou sete doses do nucleotídeo C no primeiro genitor e quatro doses no segundo genitor com oito classes genotípicas na progênie.

As linhas tracejadas representam os ângulos esperados para as ploidias estimadas, ou seja, para uma determinada ploidia, espera-se que as nuvens de ponto distribuam-se ao longo das linhas tracejadas. É importante ressaltar que tais linhas (ângulos) representam a razão esperada entre os valores das massas dos dois alelos, dependendo portanto da dose e ploidia do loco. Dessa forma, nota-se uma grande concordância entre estas linhas e as classes genotípicas obtidas. Os histogramas, localizados na parte inferior do gráfico, mostram uma forte concordância (em roxo) entre os valores das frequências genotípicas esperadas (segundo a segregação mandeliana) na progênie (em azul) e as frequências observadas (em vermelho-claro). Nota-se que no

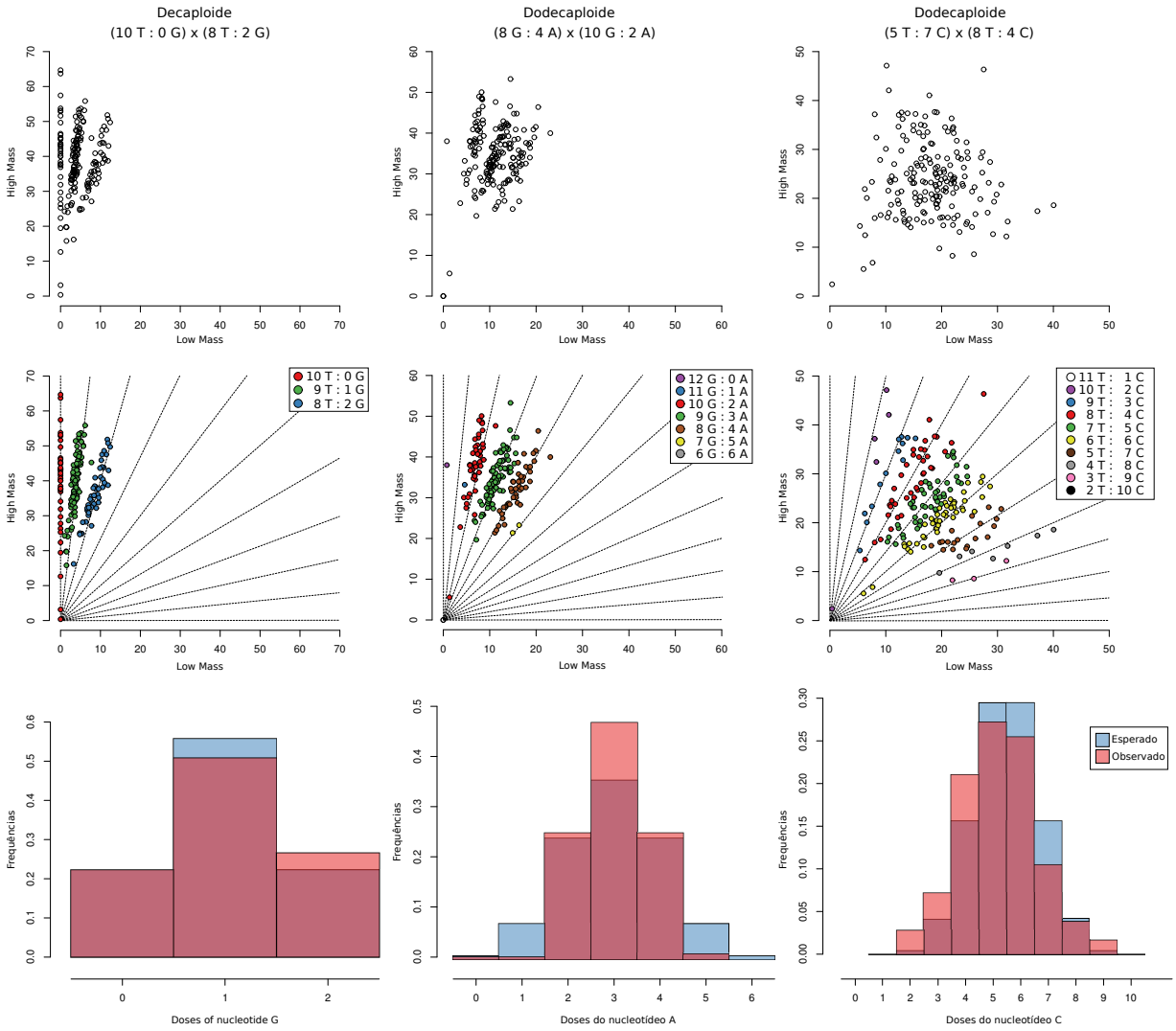

Figura 7 – Classificação de três SNPs utilizando o software SUPERMASSA. Os três gráficos superiores mostram os dados brutos. Os três gráficos intermediários mostram os genótipos alocados em diferentes classes genotípicas com os ângulos esperados para a ploidia estimada (linhas tracejadas). Os três gráficos inferiores mostram as frequências esperadas (em azul) e as frequências observadas (em vermelho-claro) na população de mapeamento. O primeiro loco, apresentado na primeira coluna, foi classificado como decaploide com dosagens zero e dois nos genitores e os genótipos da progênie foram distribuídos em três classes com frequências esperadas de $\frac{2}{9}$, $\frac{5}{9}$ e $\frac{2}{9}$, respectivamente. O segundo loco, dodecaploide com dosagens quatro e dois nos genitores e seis classes genotípicas na progênie com frequências esperadas de $\frac{5}{726}$, $\frac{26}{363}$, $\frac{8}{33}$, $\frac{130}{363}$, $\frac{8}{33}$, $\frac{26}{363}$ e $\frac{5}{726}$. O terceiro loco, também dodecaploide com dosagens sete e quatro nos genitores e oito classes genotípicas com frequências esperadas de $\frac{23}{4356}$, $\frac{185}{4356}$, $\frac{638}{4356}$, $\frac{643}{2178}$, $\frac{643}{2178}$, $\frac{638}{4356}$, $\frac{185}{4356}$ e $\frac{23}{4356}$. O fato das classes extremas serem raras nos dois últimos casos, explica porque algumas delas não foram observadas, embora o modelo leve em conta essa distribuição

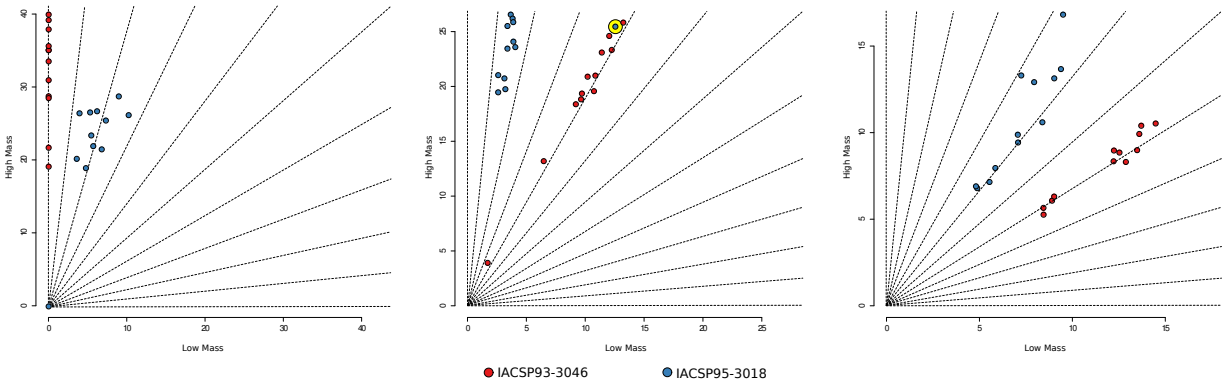

Figura 8 – Dispersão dos dados nas 12 repetições realizadas nos dois genitores. As linhas tracejadas indicam os ângulos esperados considerando-se os níveis de ploidia estimados (10, 12 e 12). O primeiro SNP foi classificado com doses zero e dois, o segundo, dois e quatro e o terceiro, sete e quatro. O círculo amarelo indica a repetição do genitor IACSP 95-3018 que se encontra fora da tendência das outras repetições

segundo SNP não foi observada uma classe genotípica (genótipo GGGGGGAAAAAA), que tem frequência esperada de 1% e no terceiro SNP não foram observadas as duas classes extremas (TTTTTTTTTTTC e TTCCCCCCCCCC), que têm frequências esperadas de 0.02%. Ainda, em todos os locos observou-se que marcadores com dosagens altas nos genitores produziram marcadores com diversas classes genotípicas na progênie, conforme esperado. A Figura 8 mostra o resultado da genotipagem dos genitores com 12 repetições para os três SNPs selecionados. Com exceção de um ponto azul (destacado pelo círculo amarelo) fora da tendência dos demais no segundo gráfico, nota-se a grande concordância entre as dosagens e níveis de ploidia estimados com os ângulos esperados (linhas tracejadas).

A Figura 9 mostra a classificação dos 241 SNPs na população de mapeamento. O gráfico à esquerda mostra uma escala de cores em função da probabilidade do SNP apresentar um certo nível de ploidia dentro do oito níveis testados. Por exemplo, a primeira linha do gráfico representa um SNP que foi classificado como hexaploide. Note que para esse SNP, a primeira coluna (referente ao nível de ploidia 6) apresenta a cor azul escura, indicando uma probabilidade maior que 0,8 do SNP pertencer ao nível de ploidia 6. Do total de 241 SNPs, 178 (74%) foram classificados com probabilidades maiores que 0,8. Nota-se ainda que para alguns SNPs essa probabilidade não foi tão alta, entretanto todos SNPs foram alocados na classe que apresentou maior probabilidade *a posteriori* segundo o método de classificação.

Seguindo esses critérios, 40 SNPs (16,6%) foram classificados como hexaploides, 32 (13,3%) como octaploides, 21 (8,7%) como decaploides, 29 (12,0%) como dodecaploides, 19 (7,9%) como tetradecaploides, 28 (11,6%) como hexadecaploides, 32 (13,3%) como octadecaploides

e 40 (16,6%) como icosaploides. Quarenta e quatro SNPs (18,3%) foram classificados como tendo no máximo uma dose em pelo menos um dos genitores. Estes SNPs estão indicados pelos pontos roxos nos gráficos central e à direita. 31 SNPs (12,7%), representados pelos pontos roxo-claros, foram classificados como tendo no máximo duas doses em pelo menos um dos genitores e 166 SNPs (69,0%), representados em laranja, foram classificados como tendo mais de duas doses em pelo menos um dos genitores.

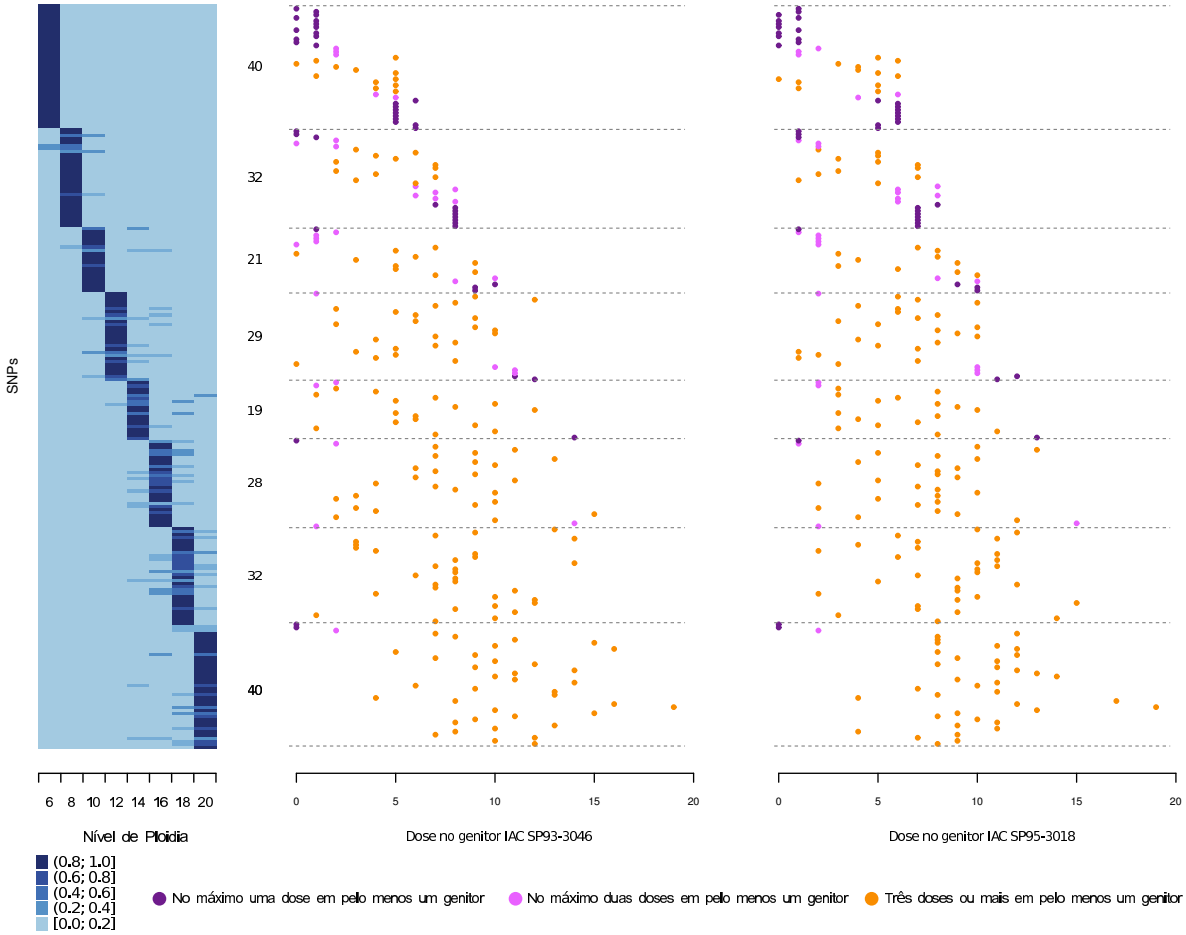

Figura 9 – Visualização da classificação de 241 SNPs na população biparental originária do cruzamento entre as variedades IACSP 93-3046 e IACSP 95-3018. O gráfico a esquerda mostra uma escala de cores em função da probabilidade do SNP apresentar o comportamento de uma determinada ploidia. Foram estudadas 8 níveis de ploidia, considerando apenas números pares (de 6 a 20). O azul escuro indica níveis de ploidia mais prováveis, enquanto que o azul claro indica níveis de ploidia menos prováveis. 40 SNPs foram classificados como hexaploides, 32 como octaploides, 21 como decaploides, 29 como dodecaploides, 19 como tetradecaploides, 28 como hexadecaploides, 32 como octadecaploides e 40 como icosaploides. Os gráficos central e a direita mostram as dosagens nas quais os marcadores dos genitores foram alocados

4.2 Mapas

Para exemplificar a aplicação do método de construção de mapas foram utilizados 40 SNPs classificados como hexaploides e 32 SNPs classificados como octaploides. É importante notar que não foram testadas ligações entre marcadores com níveis de ploidia diferentes, uma vez que não há razões biológicas para que isso ocorra.

A Figura 10 mostra os seis grupos de ligação hexaploides obtidos com o método aqui desenvolvido. Cada um dos grupos contém dois conjuntos de seis linhas verticais que representam os grupos de homologia dos genitores IACSP93-3046 e IACSP95-3018. Nestes grupos de homologia estão dispostos marcadores em diversas doses. Cada loco está representado por pontos dispostos horizontalmente. A distância entre os marcadores pode ser observada à direita dos grupos de homologia, juntamente com os nomes dos marcadores. Os marcadores SNPs estão representados por pontos coloridos, de modo que cada cor indica o nucleotídeo referente àquele polimorfismo.

Observa-se que cada loco SNP apresenta no máximo dois alelos distintos devido a sua natureza bialélica. Os marcadores AFLP e microssatélite são representados respectivamente por pontos cinza claro e pretos (presença da marca) e traços (ausência). Como os microssatélites apresentam comportamento dominante em poliploides complexos como a cana-de-açúcar, nota-se a presença de, no máximo, um polimorfismo (ponto preto) por loco por grupo de homologia. Além disso, não foram utilizados marcadores microssatélite em outras doses, pelas razões expostas na Seção 3. Da mesma maneira que os grupos hexaploides, os gráficos presentes na Figura 11 permitem a visualização de oito grupos de ligação octaploides com a disposição dos diferentes tipos de marcadores em diversas doses.

Do total de 706 marcadores (40 SNPs hexaploides, 32 SNPs octaploides, 258 AFLPs e 376 microssatélites), 76 formaram grupos de ligação hexaploides e 72 formaram grupos de ligação octaploides. O comprimento do mapa para os grupos hexaploides (Figura 10) foi de 780 cM com densidade média de $10{,}26$ cM (i.e. em média um marcador a cada $10{,}26$ cM) e para os grupos octaploides (Figura 11) o comprimento foi de $653{,}5$ cM com densidade média de $9{,}25$ cM. Os grupos de ligação hexaploides 1 e 5 apresentaram as maiores densidades de marcas, com $8{,}0$ cM e $3{,}6$ cM, respectivamente. Os demais grupos apresentaram densidades menores (grupo 2: $14{,}9$ cM; grupo 3: $9{,}17$ cM; grupo 4: $14{,}9$ cM; grupo 6: $10{,}0$ cM). Os grupos octaploides 1 e 6 apresentaram maior densidade ($6{,}40$ cM e $3{,}06$ cM, respectivamente).

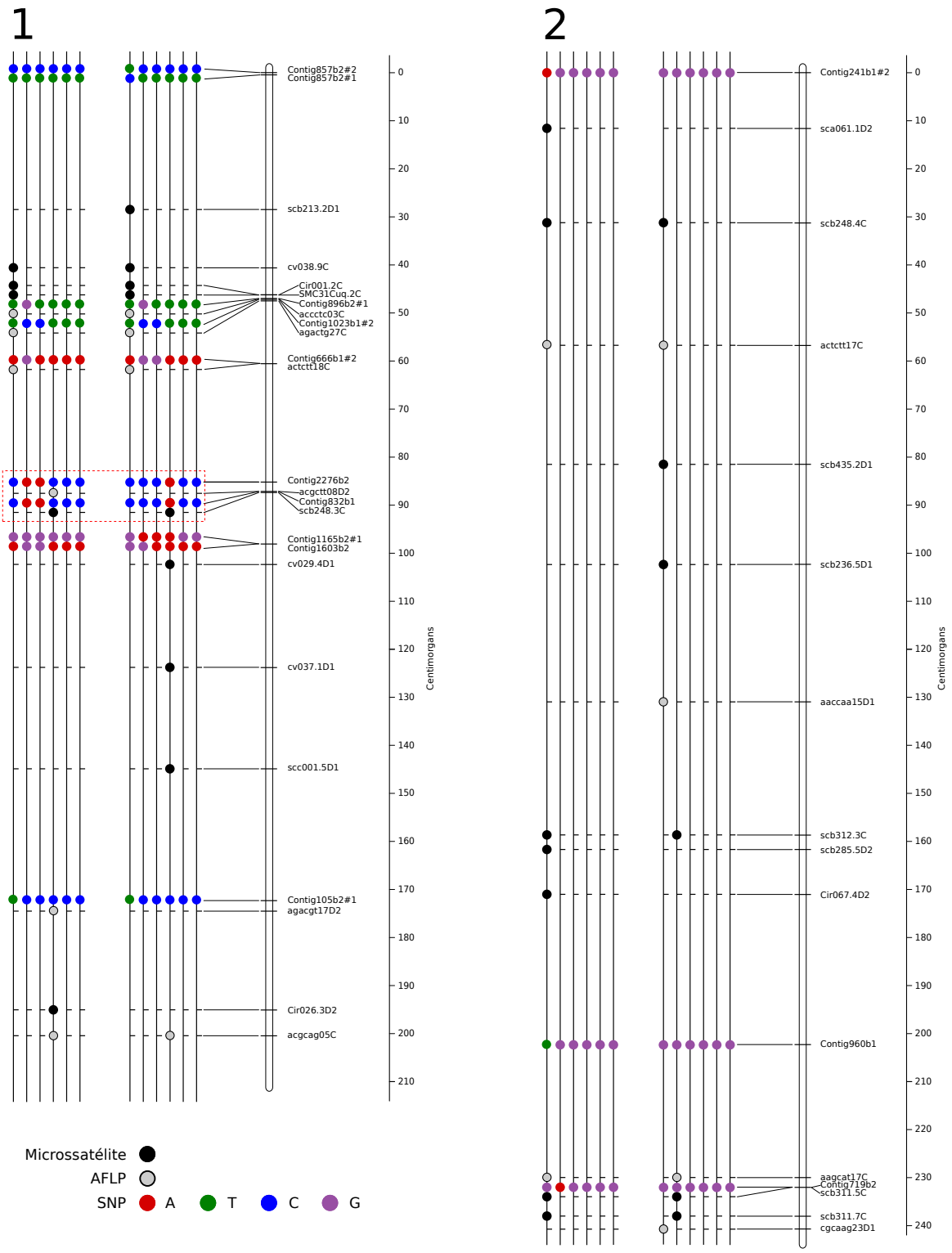

Figura 10 – Mapa genético integrado de seis grupos de homologia hexaploides contendo SNPs (representados por pontos coloridos), microssatélites (representados por pontos pretos) e AFLPs (representados por pontos cinzas). Para cada um dos grupos, os dois conjuntos de seis linhas verticais representam os grupos de homologia para os genitores IACSP 93-3046 e IACSP 95-3018. Cada loco (pontos dispostos horizontalmente) pode apresentar até dois alelos. A dosagem do marcador e o nucleotídeo referente a ela pode ser verificada pela diferença de cores dentro do mesmo loco. Por exemplo, o marcador *Contig1603b2* apresenta duas doses do nucleotídeo G e quatro doses do nucleotídeo A em ambos os genitores. Ao lado dos grupos, as distâncias estão representadas juntamente com os nomes dos marcadores

(continua)

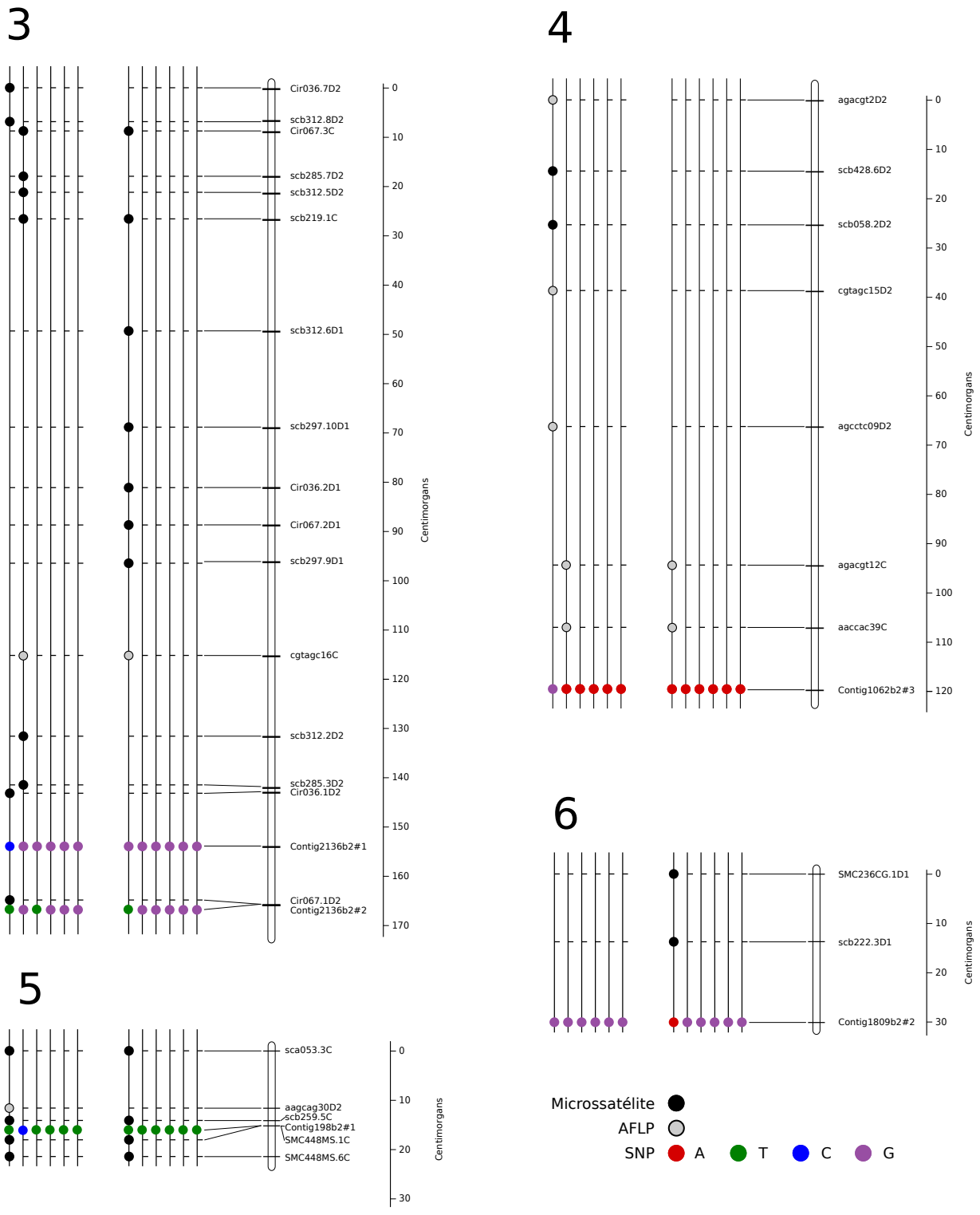

Figura 10 – Mapa genético integrado de seis grupos de homologia hexaploides contendo SNPs (representados por pontos coloridos), microssatélites (representados por pontos pretos) e AFLPs (representados por pontos cinzas). Para cada um dos grupos, os dois conjuntos de seis linhas verticais representam os grupos de homologia para os genitores IACSP 93-3046 e IACSP 95-3018. Cada loco (pontos dispostos horizontalmente) pode apresentar até dois alelos. A dosagem do marcador e o nucleotídeo referente a ela pode ser verificada pela diferença de cores dentro do mesmo loco. Por exemplo, o marcador *Contig1603b2* apresenta duas doses do nucleotídeo G e quatro doses do nucleotídeo A em ambos os genitores. Ao lado dos grupos, as distâncias estão representadas juntamente com os nomes dos marcadores (conclusão)

Os demais grupos octaploides apresentaram densidades de $8{,}17$ cM (grupo 2), $8{,}79$ cM (grupo 3), $11{,}32$ cM (grupo 4), $12{,}70$ cM (grupo 5), $14{,}46$ cM (grupo 7) e $8{,}85$ cM (grupo 8). Tanto o grupo 1 hexaploide quanto o grupo 1 octaploide serão aqui discutidos por apresentarem mais marcadores, maior densidade e situações interessantes que exemplificam as vantagens do novo método aqui desenvolvido.

Nas Figuras 10 e 11, nota-se que além das distâncias entre os marcadores, comumente apresentadas em mapas genéticos, é possível observar a fase de ligação entre eles. Com essa informação foi possível obter os haplótipos de cada cromossomo homólogo de ambos genitores. No caso dos grupos de homologia hexaploides, foram obtidos 12 haplótipos para cada grupo (seis em cada genitor) e no caso dos octaploides 16 haplótipos. Para exemplificar a disposição dos marcadores nos grupos de homologia, considere os marcadores em destaque (retângulo tracejado) no grupo de ligação 1 da Figura 10. O marcador *Contig2276b2* possui duas doses do nucleotídeo `A` no primeiro genitor e uma dose no segundo genitor. Tratando-se de um grupo hexaploide e sendo o outro nucleotídeo nesse loco uma citosina (`C`), esse marcador apresenta quatro doses de `C` no primeiro genitor e cinco doses no segundo genitor. A $1,94$ cM desse marcador encontra-se a marca AFLP *acgctt08D2*, que apresenta polimorfismo apenas no primeiro genitor. Ainda, pode-se dizer que o polimorfismo do AFLP está presente em um dos homólogos que apresentou o nucleotídeo `C` no marcador anterior. O próximo marcador, *Contig832b1*, a zero cM do marcador anterior, tem exatamente a mesma disposição nos homólogos que o marcador *Contig2276b2*, ou seja, os nucleotídeos `A` encontram-se nos mesmos cromossomos homólogos, assim como os nucleotídeos `C`. O próximo marcador, microssatélite *scb248.3C*, também a zero cM do marcador anterior, apresenta polimorfismos nos dois genitores, ligados a nucleotídeos `A` e `C`. Dessa forma, pode-se verificar todos os arranjos que esses marcadores podem se encontrar em genomas complexos. É evidente que tal representação é muito mais informativa do que as usualmente apresentadas para poliploides.

No grupo 1 octaploide (Figura 11) pode-se observar uma situação bastante complicada que exemplifica a potencialidade do método (retângulo tracejado). Observa-se um marcador AFLP (*cgtaca29C*) em fase de ligação com um marcador microssatélite (*scb082.1C*), que por sua vez está ligado a um marcador SNP (*Contig1062b2#4*) com duas doses do nucleotídeo `C` no primeiro genitor e três doses no segundo genitor. Este último encontra-se ligado a outros SNPs (*Contig1929b2* e *Contig1055b1*). A maneira com que estes marcadores encontram-se ligados não é trivial, podendo-se observar a formação de seis haplótipos distintos para o primeiro genitor: `Af-Ms-T-G-T`, `o-o-C-A-C`, `o-o-C-G-C`, `o-o-T-A-C`, `o-o-T-A-T` e

o-o-T-G-T, sendo kAf o polimorfismo do marcador AFLP, Ms o polimorfismo do microssatélite, o a ausência da marca e as demais letras polimorfismos dos SNPs. No segundo genitor observa-se dois haplótipos diferentes: Af-Ms-T-A-T e o-o-C-A-T. Evidentemente, essa análise de haplótipos pode ser estendida para o cromossomo homólogo inteiro, podendo então caracterizá-lo. Dessa forma, nota-se que dos oito cromossomos homólogos presentes no primeiro genitor, apenas dois deles apresentaram haplótipos iguais, ocorrendo o mesmo no segundo genitor.

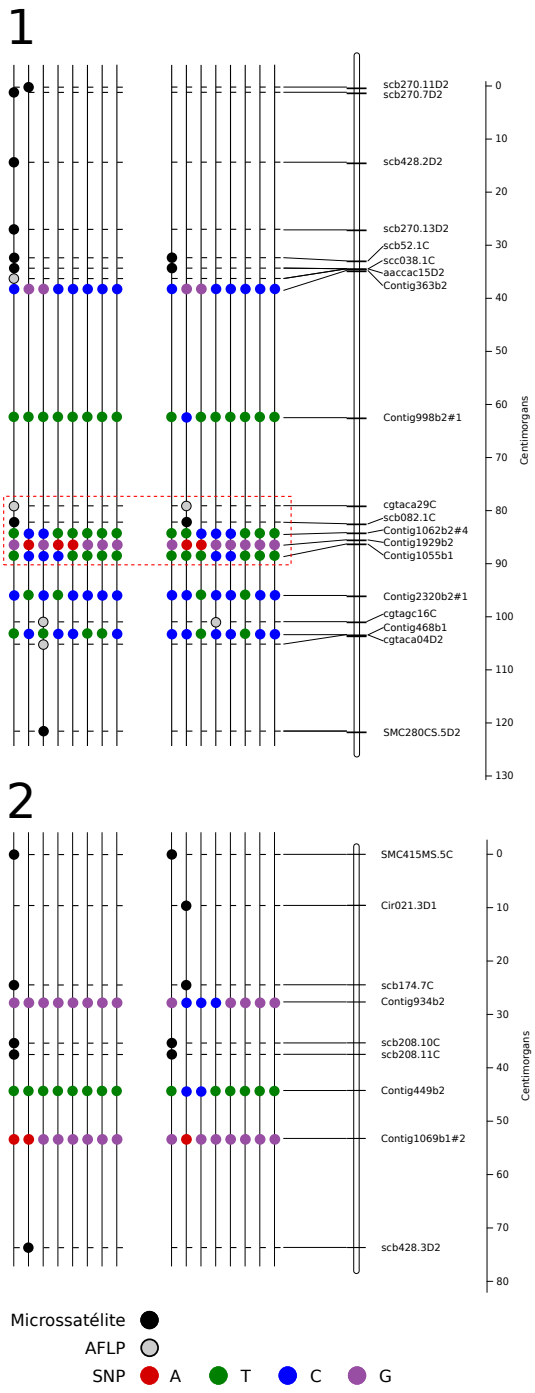

Figura 11 – Mapa genético integrado de oito grupos de homologia octaploides contendo SNPs (representados por pontos coloridos), microssatélites (representados por pontos pretos) e AFLPs (representados por pontos cinzas). Para cada um dos grupos, os dois conjuntos de oito linhas verticais representam os grupos de homologia para os genitores IACSP 93-3046 e IACSP 95-3018. Cada loco (pontos dispostos horizontalmente) pode apresentar até dois alelos. A dosagem do marcador e o nucleotídeo referente a ela pode ser verificada pela diferença de cores dentro do mesmo loco. Ao lado dos grupos, as distâncias estão representadas juntamente com os nomes dos marcadores (continua)

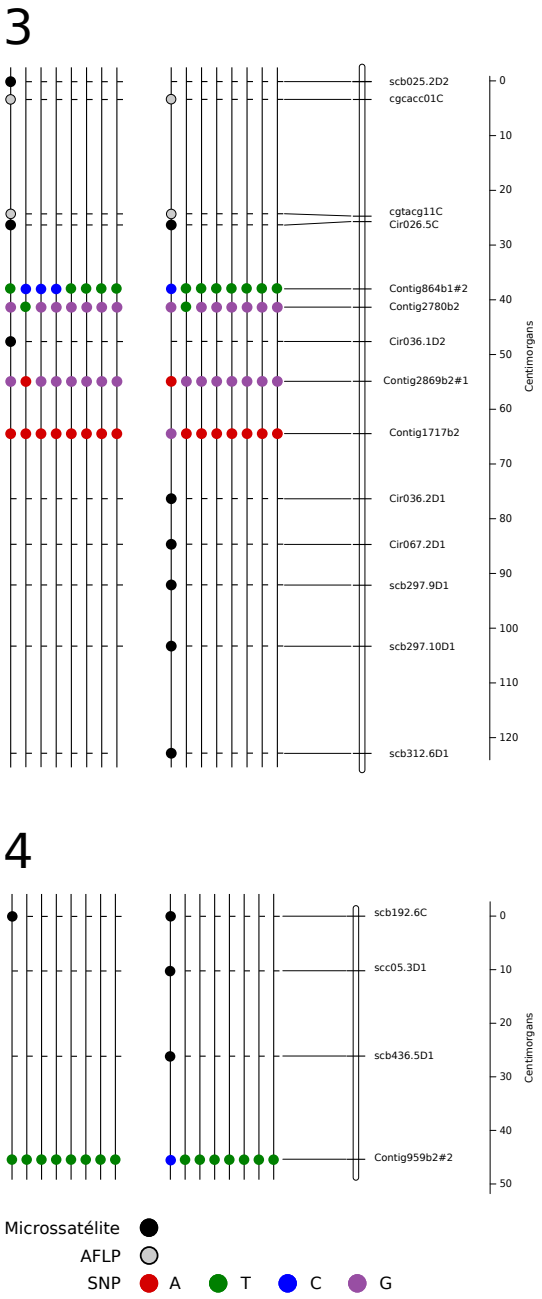

Figura 11 – Mapa genético integrado de oito grupos de homologia octaploides contendo SNPs (representados por pontos coloridos), microssatélites (representados por pontos pretos) e AFLPs (representados por pontos cinzas). Para cada um dos grupos, os dois conjuntos de oito linhas verticais representam os grupos de homologia para os genitores IACSP 93-3046 e IACSP 95-3018. Cada loco (pontos dispostos horizontalmente) pode apresentar até dois alelos. A dosagem do marcador e o nucleotídeo referente a ela pode ser verificada pela diferença de cores dentro do mesmo loco. Ao lado dos grupos, as distâncias estão representadas juntamente com os nomes dos marcadores (continuação)

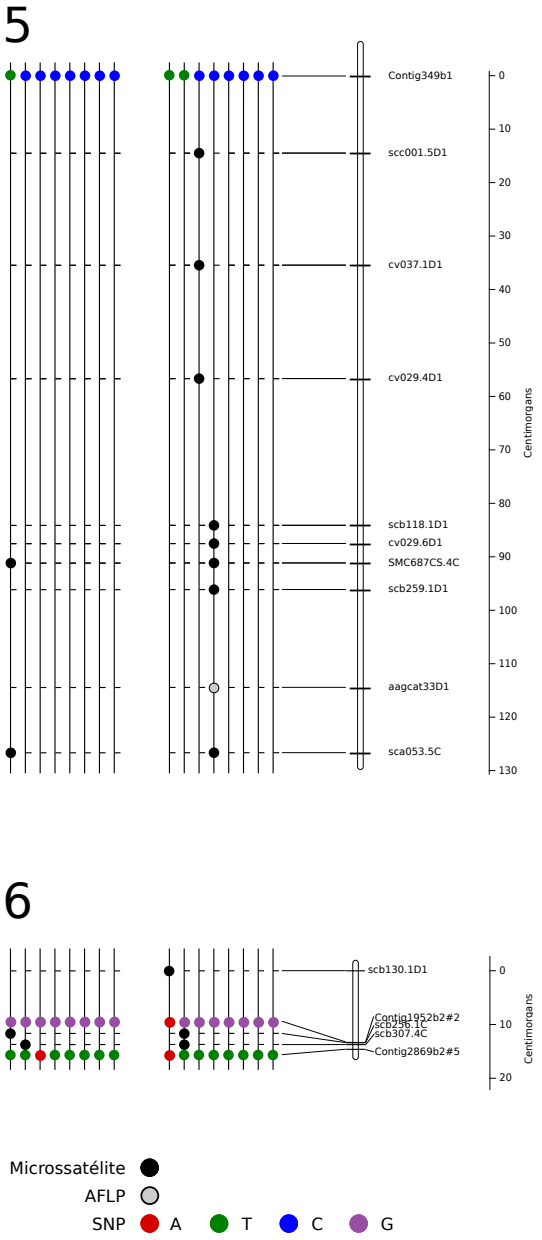

Figura 11 – Mapa genético integrado de oito grupos de homologia octaploides contendo SNPs (representados por pontos coloridos), microssatélites (representados por pontos pretos) e AFLPs (representados por pontos cinzas). Para cada um dos grupos, os dois conjuntos de oito linhas verticais representam os grupos de homologia para os genitores IACSP 93-3046 e IACSP 95-3018. Cada loco (pontos dispostos horizontalmente) pode apresentar até dois alelos. A dosagem do marcador e o nucleotídeo referente a ela pode ser verificada pela diferença de cores dentro do mesmo loco. Ao lado dos grupos, as distâncias estão representadas juntamente com os nomes dos marcadores (continuação)

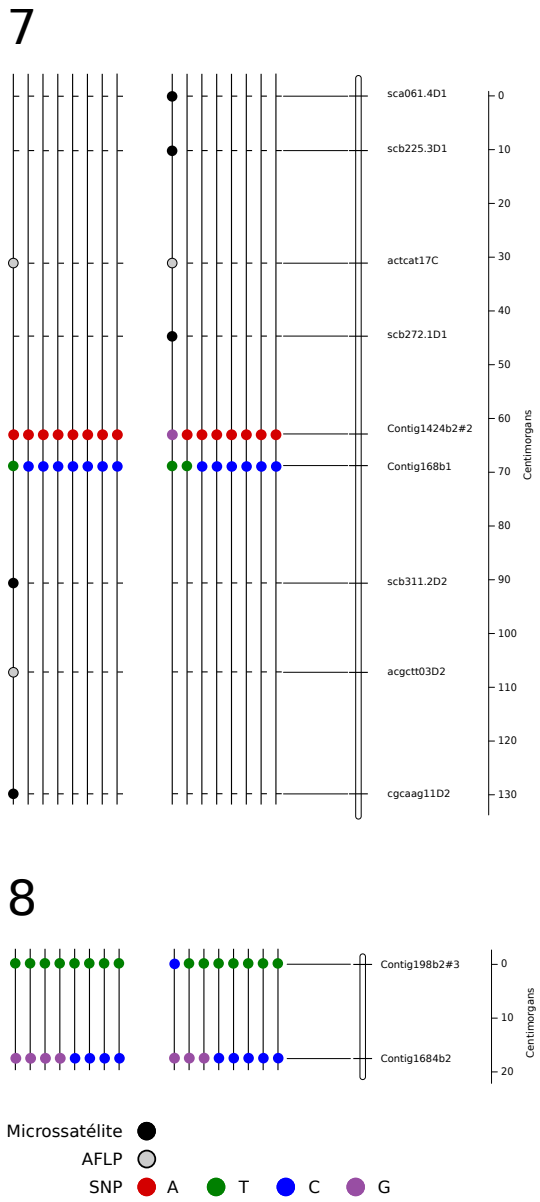

Figura 11 – Mapa genético integrado de oito grupos de homologia octaploides contendo SNPs (representados por pontos coloridos), microssatélites (representados por pontos pretos) e AFLPs (representados por pontos cinzas). Para cada um dos grupos, os dois conjuntos de oito linhas verticais representam os grupos de homologia para os genitores IACSP 93-3046 e IACSP 95-3018. Cada loco (pontos dispostos horizontalmente) pode apresentar até dois alelos. A dosagem do marcador e o nucleotídeo referente a ela pode ser verificada pela diferença de cores dentro do mesmo loco. Ao lado dos grupos, as distâncias estão representadas juntamente com os nomes dos marcadores (conclusão)

5 DISCUSSÃO

Em essência, todos os mapas genéticos de autopoliploides (com exceção dos autotetraploides) têm sido construídos usando marcadores com comportamento dominante, os quais segregam em apenas duas classes (presença/ausência) em uma população de mapeamento. Isso decorre do fato dos sistemas de marcadores utilizados até então não permitirem a separação de genótipos com uma ou mais cópias do mesmo alelo. De forma a conseguir utilizar os programas de mapeamento (softwares) desenvolvidos para espécies diploides, uma simplificação do problema foi proposta (WU et al., 1992). Como marcadores em dose única sempre segregam na proporção mendeliana 1:1, independente do nível de ploidia da espécie, a seleção de locos com esse comportamento permite a construção de mapas de forma análoga àquela usada em retrocruzamentos.

No caso da cana-de-açúcar, as populações de mapeamento utilizadas inicialmente foram delineadas de forma a medir polimorfismo em apenas um dos genitores. O único problema para a análise dos dados nessa abordagem é o desconhecimento da fase de ligação entre os locos. Porém, dentro desse mesmo período histórico (década de 1990), o problema foi solucionado em diploides com o desenvolvimento da técnica chamada *pseudo-testcross*, específica para marcadores com segregação 1:1 (GRATTAPAGLIA; SEDEROFF, 1994). Nessa abordagem, os dados são codificados de forma especial nos programas computacionais e um mapa é construído para cada genitor. A quase totalidade dos mapas de cana-de-açúcar desenvolvidos até o momento fez uso dessa abordagem. Porém, no caso das espécies diploides, desenvolvimentos teóricos posteriores possibilitaram a construção de mapas com marcadores com todos tipos de segregação (1:1:1:1, 1:2:1, 3:1 e 1:1). Essas segregações puderam ser observadas em função do surgimento de novos tipos de marcadores moleculares, tais como microssatélites e SNPs. É importante ressaltar que o *pseudo-testcross* foi sugerido em função da disponibilidades na época do marcador RAPD (*Random Amplified Polymorphic DNA* - WILLIAMS et al., 1990) que, por ser dominante, possui segregação 1:1.

Apesar dos avanços tecnológicos em diploides permitirem a utilização simultânea de locos com diferentes padrões de segregação, pesquisas desenvolvidas pelo grupo a que pertencem os envolvidos com a presente tese permitiram observar que o método de análise tradicional, baseado em análises de dois e três pontos, poderia ser sensivelmente melhorado com o uso do

modelo de Markov oculto. Tal modelo permite a utilização simultânea da informação de todos os marcadores do mesmo grupo de ligação. Isso motivou a criação do software OneMap (MARGARIDO; SOUZA; GARCIA, 2007; MARGARIDO; MOLLINARI; GARCIA, 2011). É importante mencionar que não só os diplóides se beneficiaram dessa abordagem, já que marcadores em dose única em autopoliplóides, quando presentes em ambos os genitores, segregam na proporção 3:1 e isso permite a construção de mapas integrados também em cana-de-açúcar, análogo ao que ocorre em diplóides (GARCIA et al., 2006; OLIVEIRA et al., 2007; PASTINA et al., 2012).

Embora a construção de mapas integrados implique num avanço conceitual no mapeamento genético, vale mencionar que no caso dos poliploides, os marcadores com segregação 3:1, usualmente empregados na integração, não são tão informativos quanto aqueles com outros padrões de segregação, como 1:1:1:1 e 1:2:1. Para diversos conjuntos de dados que foram analisados pelo grupo de pesquisa, tal integração não foi possível, ou mesmo recomendada. Porém, é importante ressaltar que essa limitação não ocorre devido à abordagem em sim, mas devido ao tipo de dados disponível. Deve haver uma quantidade suficiente de locos com polimorfismos nos dois genitores para que a integração seja feita de maneira apropriada. Entende-se que a integração deve ser buscada sempre que possível, dada suas vantagens; isso foi um dos motivadores para o presente trabalho.

Uma primeira abordagem que poderia ser investigada para autopoliploides seria a inclusão de locos com outros padrões de segregação e outras doses (além de dose única) no mapa. Outros tipos de segregação já foram investigados em cana-de-açúcar, mas sempre considerando polimorfismo em apenas um dos genitores e marcadores com comportamento dominante, tais como marcadores com segregação 11:3, 13:1 e 69:1 em auto-octaploides. Embora aumentem a cobertura do genoma por amostrarem regiões que não exclusivamente aquelas em dose única, é fácil notar que esses marcadores são pouco informativos devido ao seu comportamento dominante. Por exemplo, um loco com dose tripla em apenas um dos genitores em um cruzamento auto-octaploide tem a segregação 13:1. Neste caso espera-se que apenas 7% dos indivíduos da população segregante não apresentem banda (ou ausência do alelo). É evidente que indivíduos sem a presença de bandas poderiam não estar presentes em populações de tamanho reduzido levando à conclusão errônea de que o marcador em questão é monomórfico. A expansão desse conceito para incluir combinações de marcadores com outras doses e com polimorfismos em ambos os genitores (e que portanto, permitiriam a integração dos mapas) é trivial e poderia ser considerada, mas o problema da baixa informatividade dos marcadores persistiria. Não pode

deixar de ser mencionado o fato de que as segregações de marcadores com mais de uma dose só podem ser deduzidas se o nível de ploidia da espécie for conhecido, o que nem sempre é o caso. Conforme observado nos resultados do presente trabalho, existe uma variação do nível de ploidia dentro do genoma da cana-de-açúcar e portanto o padrão de segregação esperado não é conhecido *a priori* para cada loco.

Neste cenário insere-se o projeto temático do qual faz parte essa tese. Foi proposto o uso de uma nova classe de marcadores, os SNPs. Além de sua grande abundância no genoma, os SNPs também permitem a leitura do nível de dosagem de cada um dos locos por se tratar de uma técnica quantitativa. Nessa técnica, por exemplo, um indivíduo tetraploide com três doses ($AAAa$) terá maior intensidade para o alelo A e menor intensidade para o alelo a que um indivíduo com genótipo $AAaa$. O fato dessa técnica permitir a leitura da intensidade dos dois sinais (referentes aos alelos A e a), torna o comportamento do marcador codominante e possibilita estimar a ploidia e a dose de cada loco.

Após a obtenção dos dados, o primeiro desafio enfrentado foi classificar cada um dos indivíduos numa dada classe genotípica. No exemplo acima, os indivíduos com 3 alelos a formam um categoria, aqueles com 2 alelos a formam uma outra categoria, e assim por diante. A forma tradicional de se fazer isso depende do conhecimento do número de grupos antecipadamente, o que acontece para espécies diplóides (3 classes usualmente estudadas na genética mendeliana: AA, Aa e aa). No entanto, para os poliploides com o nível de ploidia desconhecido isso não ocorre. O novo método proposto e implementado no software SUPERMASSA realiza essa tarefa e pode ser descrito em duas etapas. Inicialmente, modela-se a probabilidade de que um indivíduo com um genótipo conhecido tenha gerado a intensidade observada pelo equipamento usado na genotipagem (assumindo que a intensidade observada seja proporcional à dosagem do alelo correspondente). Numa etapa seguinte, calcula-se a probabilidade de que o conjunto de genótipos observados tenha ocorrido, dada a estrutura populacional em questão. Isto permite que o nível de ploidia também possa ser estimado, o que é crucial no caso da cana-de-açúcar e de vários outros poliploides. Além de estimar a ploidia, a classificação obtida com o método mostra o comportamento codominante do marcador. Por exemplo, a segregação 13:1 descrita anteriormente, se observada usando-se um marcador SNP, será 1:6:6:1 com evidente aumento no poder de resolução. Só foi possível aproveitar essa vantagem graças a correta escolha da tecnologia utilizada para a realização da genotipagem com alto grau de precisão (espectrometria de massa). Fica claro que diversas novas possibilidades para estudos genéticos foram abertas graças a esse grande avanço tecnológico e metodológico.

Estudos citológicos feitos em cana-de-açúcar têm demonstrado o alto nível de ploidia e aneuploidia presentes no gênero *Saccharum*. Os cultivares modernos de cana-de-açúcar são derivados de cruzamentos realizados há cerca de um século entre a espécie domesticada *S. officinarum* ($x = 10$, $2n = 8x = 80$, sendo x o número básico de cromossomos) e a espécie selvagem *S. spontaneum* ($x = 8$, $2n = 5 - 16x = 40 - 128$). Após esse cruzamento interespecífico, uma série de retrocruzamentos com *S. officinarum* foi realizada. Esse processo, chamado de *nobilização*, teve como objetivo recuperar características agronômicas de interesse presentes em *S. officinarum* e diluir as características desfavoráveis presentes em *S. spontaneum* (**??**). Em seu famoso trabalho publicado em 1996, D'Hont et al., usando técnicas de citogenética molecular, demonstraram que cerca de 80% dos cromossomos da cultivar francesa R570 são originários de *S. officinarum*, 10% de *S. spontaneum* e 10% são produto de recombinação entre essas duas espécies. Proporções semelhantes também foram observadas em outros estudos (CUADRADO et al., 2004; D'HONT, 2005; PIPERIDIS; PIPERIDIS; D'HONT, 2010). Entretanto, esses resultados foram obtidos utilizando-se poucos cultivares, sendo então muito difícil extrapolá-los para a cana-de-açúcar de uma maneira geral. **??**) argumentam que, devido ao seu alto nível de ploidia e similaridades entre os cromossomos de *S. officinarum* e *S. spontaneum*, informações como o número básico de cromossomos, bem como o comportamento dos cromossomos selvagens durante a introgressão e sua exata contribuição para os cultivares modernos, não podem ser definitivamente estabelecidas. Nesse contexto, os níveis de ploidia encontrados nesse trabalho, variando de 6 a 20, estão condizentes com as informações encontradas na literatura. Ainda, a classificação dos SNPs mostra não haver o predomínio de nenhum nível de ploidia dentro do genoma. Diferentemente dos mapas genéticos usualmente publicados, o método desenvolvido no presente trabalho permite o estudo de diferentes níveis de ploidia dentro do genoma, que é exatamente o caso da cana-de-açúcar e foi mostrado por D'Hont et al. (1996) e D'Hont (2005) através de estudos citológicos. Entretanto, com base nos resultados aqui apresentados, nada pode se afirmar a respeito da natureza interespecífica da cana-de-açúcar.

Ao contrário do que tem sido relatado na literatura, como por exemplo nos trabalhos de (AITKEN; JACKSON; MCINTYRE, 2005; EDME; GLYNN; COMSTOCK, 2006; AITKEN; JACKSON; MCINTYRE, 2007; ALWALA et al., 2008; BAKER; JACKSON; AITKEN, 2010; HELLER-USZYNSKA et al., 2010) os resultados aqui apresentados mostram que não há o predomínio de marcadores em dose única nos genomas dos cultivares utilizados no cruzamento biparental. Acredita-se que essa discrepância de resultados deve-se principalmente ao fato dos marcadores utilizados nesses trabalhos terem comportamento dominante. Para que um marca-

dor seja usado em uma população biparental, geralmente é feito um teste de segregação, ou seja, uma amostra da população é genotipada à procura de polimorfismos. Caso o marcador apresente polimorfismo na amostra considerada, ele é utilizado na população inteira. Entretanto, essa amostra geralmente é muito pequena (de 15 a 20 indivíduos) se comparada aos números necessários para que seja encontrado um polimorfismo com marcadores com mais de uma dose. Conforme o exemplo apresentado acima, um marcador em dose tripla em um auto-octaploide não apresentará a banda em apenas 7% da população segregante. Devido ao pequeno tamanho da amostra, muitas vezes esses marcadores são considerados monomórficos, levando a conclusões errôneas sobre a presença de polimorfismos na população segregante. Outro problema encontrado é que marcadores que apresentarem dosagens maiores do que a metade do nível de ploidia sempre apresentarão presença de banda na progênie, uma vez que, independentemente dos cromossomos que compuserem o gameta, pelo menos um deles apresentará uma dose. Locos com tal característica também serão classificados como monomórficos quando na verdade apenas constituem um loco com alta dose. Dessa forma, é muito provável que a afirmação de que os marcadores em dose única são os mais importantes para o mapeamento está baseada em limitações técnicas e não condizem com a realidade da grande maioria das espécies poliploides, incluindo a cana-de-açúcar.

Ainda no contexto de dosagem alélica, pode-se especular que polimorfismos em dose única num autopoliploide estável ocorre devido a mutações pontuais depois do processo de poliploidização, uma vez que se tivessem ocorrido antes desse processo as doses seriam maiores. Todos os trabalhos citados (AITKEN; JACKSON; MCINTYRE, 2005; EDME; GLYNN; COMSTOCK, 2006; AITKEN; JACKSON; MCINTYRE, 2007; ALWALA et al., 2008; BAKER; JACKSON; AITKEN, 2010; HELLER-USZYNSKA et al., 2010) envolvem o cruzamento de *S. officinarum* e algum cultivar comercial ou *S. spontaneum* e muitas vezes o polimorfismo é medido apenas em *S. officinarum*. Além desses polimorfismos serem medidos usando-se marcadores em dose única, os quais possuem um viés intrínseco, não é recomendável generalizar essa afirmação para cultivares comerciais. Num cruzamento interespecífico, haverá uma perturbação no genoma e também no comportamento meiótico dos descendentes, os quais sofrerão mudanças no nível de ploidia. Quando isso ocorre, o resultado pode ser um múltiplo da geração anterior, uma vez que o comportamento meiótico pode ter sido afetado com a fusão de genomas interespecíficos. Logo, pode-se especular que o aumento do nível de ploidia acarretaria o aumento da dosagem dos marcadores, fenômeno esse que foi observado nos dados aqui apresentados (Figura 9). É evidente que mais estudos são necessários para validação dessa teoria, como por exemplo, a

construção de mapas em populações derivadas da autofecundação de *S. officinarum* e estudos citogenéticos de comportamento meiótico.

O método de mapeamento aqui desenvolvido mostrou-se eficiente para construir mapas em espécies poliploides quando os marcadores têm comportamento dominante; no caso, SNPs. Isso é um grande avanço em relação aos métodos desenvolvidos para autotetraploides, que já se beneficiam da codominância e do multialelismo dos marcadores moleculares desde o começo dos anos 2000 (LUO et al., 2000, 2001; LUO; ZHANG; KEARSEY, 2004; LUO et al., 2006; LEACH et al., 2010). Autotetraploides são, inquestionavelmente, os autopoliploides que possuem a maior quantidade de informações acumuladas na literatura. Como exemplo, podemos citar a grande quantidade de informações que existem a respeito do comportamento meióticos de autotetraploides, como formação de multivalentes e dupla redução (BURNHAM, 1962; SYBENGA, 1972; APPELS et al., 1998; SINGH, 2003). Dessa forma, toda teoria genética-estatística desenvolvida para autotetraploides pode lançar mão dessa grande quantidade de informações, como feito por Leach et al. (2010). No presente trabalho, foram utilizados marcadores codominantes e bialélicos (SNPs) e foi necessário o desenvolvimento de um método que levasse em conta essa característica para qualquer nível de ploidia e dosagem. Isto foi possível graças a Equação 8 que resume qualquer caso possível e torna viável o cálculo de qualquer probabilidade de transição necessária para a construção do mapa. Este fato foi exemplificado com dois níveis de ploidia: $m = 6$ e $m = 8$.

Nos mapas aqui apresentados puderam ser observados diversos padrões complexos de fases de ligação. Em algumas situações isso permitiu discriminar cromossomos homólogos dentro de um mesmo grupo de homologia graças aos haplótipos formados pela combinação dos alelos presentes nos SNPs, AFLPs e microssatélites. Mesmo sendo usados apenas 241 SNPs, dos quais 40 apresentaram comportamento hexaploide e 32 apresentaram comportamento octaploide, essa discriminação pode ser feita dentro desses níveis de ploidia. Embora fossem esperados poucos grupos de homologia dentro de cada nível de ploidia, como mostrado por D'Hont et al. (1996) e D'Hont (2005), notam-se seis grupos para $m = 6$ (Figura 10) e oito grupos para $m = 8$ (Figura 11). Provavelmente isso deve-se ao número reduzido de marcadores SNPs utilizados para a construção do mapa. Espera-se que com a continuação do projeto, esses números aumentem e possibilitem a ligação dos grupos aqui mostrados. Luo et al. (2001) apresentaram uma representação de um mapa simulado similar à representação utilizada no presente trabalho. Entretanto, não é possível fazer comparações do mapa aqui obtido com os mapas existentes na literatura de cana-de-açúcar, já que é a primeira vez que esse tipo de representação é usada nessa espécie ou

mesmo em espécies com níveis de ploidias altos. Acredita-se que o mapa aqui apresentado corresponde ao que deveria ser buscado no mapeamento de todas espécies autopoliploides. Ainda, é importante notar que ligações entre AFLPs e microssatélites foram possíveis pela presença dos SNPs, os quais funcionaram como "marcadores ponte".

Com a disponibilidade cada vez maior de marcadores SNPs em cana-de-açúcar e em outras espécies autopoliploides com altos níveis de ploidia, algumas questões biológicas poderão ser formuladas produzindo diferentes matrizes de transição $\mathbf{T}$ na Equação 9. Isso é particularmente importante em genitores com genótipos ancestrais diferentes ou mesmo cruzamentos interespecíficos. Alguns trabalhos apontam a existência de pareamento preferencial em cana-de-açúcar (HOARAU et al., 2001; JANNOO et al., 2004) e em autopoliploides de maneira geral (CAO; OSBORN; DOERGE, 2004). Este fenômeno pode ser incluído no modelo atribuindo-se alguma distribuição de probabilidade não uniforme para Ψ. Essa modificação poderia ser levada em conta na Equação 7. Ainda, segundo Guimarães, Sills e Sobral (1997), pareamento preferencial pode indicar estágios intermediários do processo conhecido como *diploidização*, no qual os rearranjos cromossômicos e o silenciamento gênico ocorre de maneira extensiva em um genoma poliploide que a espécie assume o comportamento diploide (SOLTIS; SOLTIS, 1999; WOLFE, 2001; COMAI, 2005).

Vale ressaltar que, ao contrario de diversos trabalhos de mapeamento conduzidos em autotetraploides (WU et al., 2001; LUO et al., 2006; LEACH et al., 2010), no presente trabalho não foi investigado o fenômeno da dupla redução. Esse fenômeno não foi considerado uma vez que não há evidências citogenéticas de sua ocorrência em autopoliploides com altos níveis de ploidia. A dupla redução ocorre quando cromátides irmãs ou partes delas migram para o mesmo gameta e geralmente é observada quando ocorre a formação de multivalentes. D'Hont et al. (1998) e Piperidis, Piperidis e D'Hont (2010) que apontam a grande maioria dos cromossomos em cana-de-açúcar formam bivalentes. Bielig, Mariani e Berding (2003) estudou a microesporogênese em um cultivar comercial de cana-de-açúcar e observou o predomínio da formação de bivalentes. Raramente foram observados univalentes e configurações complexas como tetravalentes não foram observadas. Nesse contexto, como a formação de multivalentes é rara, podemos concluir que também é raro o fenômeno da dupla redução em poliploides com altos níveis de ploidia. Dessa forma, podemos dizer que as pressuposições de pareamento bivalente e não ocorrência de dupla redução não comprometem o bom desempenho do modelo, uma vez que ambas têm grande fundamento biológico. De qualquer forma, é perfeitamente possível estender o modelo aqui apresentado caso surjam evidências biológicas a respeito desses fenômenos.

Os novos mapas genéticos que serão construídos com a metodologia aqui proposta permitirão grandes avanços nos estudos genéticos de poliploides. Uma vez que o mapa multiponto usando todas as doses possíveis tenha sido construído, será possível desenvolver modelos para o mapeamento de QTLs levando em conta o efeito de dosagem. A inclusão de múltiplas doses no mapeamento genético de poliploides é extremamente importante, pois possibilita o estudo do efeito da dosagem alélica no mapeamento de QTLs (DOERGE; CRAIG, 2000; CAO; CRAIG; DOERGE, 2005). Esses efeitos de dosagem foram explorados por diversos autores mostrando diferentes padrões de expressão gênica em função da dosagem alélica e nível de ploidia. Como exemplo, podemos citar Guo, Davis e Birchler (1996) em milho, Galitski et al. (1999) em *Saccharomyces cerevisiae* e Wang et al. (2006) em Arabidopsis. Osborn et al. (2003) apresentam uma revisão sobre o efeito da dosagem alélica na expressão gênica em especies poliploides. O estudo mostra que os efeitos de dosagem alélica são geralmente observados em genótipos heterozigóticos como níveis de expressão gênica intermediárias aos extremos homozigóticos. Entretanto, este fenômeno não tende a expandir a amplitude dos fenótipos, mas sim produzir classes fenotípicas intermediárias com vantagens seletivas. Esses estudos mostram que a inclusão de múltiplas doses no mapeamento genético e consequentemente no mapeamento de QTLs em espécies autopoliploides é de fundamental importância. Ainda, a metodologia aqui apresentada permitirá o estudo de sintenia entre genótipos de cana-de-açúcar, incluindo ancestrais de cruzamentos interespecíficos, e outras espécies com ancestrais comuns à cana-de-açúcar, como é o caso do sorgo e do arroz. É importante mencionar que as ideias apresentadas aqui, servirão como base para o estudo de transmissão de haplótipos em poliploides, fenômeno bastante conhecido em diploides, mas pouco estudado em poliploides.

Estima-se que sejam necessários dois marcadores a cada um milhão de bases para auxiliar na montagem do genoma. Isso implica em cerca de 20000 marcadores em dose única distribuído igualmente entre os cerca de 115 cromossomos dos cultivares de cana-de-açúcar (WANG et al., 2010). Entretanto, os mapas atuais cobrem apenas uma porção do genoma, com não mais de 2000 marcadores. A estratégia aqui apresentada permite que a cobertura necessária seja obtida com cerca de 3000 marcadores considerando-se todas as doses possíveis. Como há a possibilidade de detecção de recombinações em mais de um bivalente, a população experimental necessária para tanto não seria maior do que as que são viáveis em pesquisas de mapeamento (cerca de 300 ou 400 indivíduos). Ao contrário do que parece estabelecido nos trabalhos que utilizaram marcadores com comportamento dominante, o presente trabalho mostrou que marcadores em doses altas são os mais informativos, pois proporcionam mais oportunidades para

detecção de eventos de recombinação.

6 CONCLUSÃO

O método aqui apresentado mostrou-se adequado para a construção de mapas genéticos em autopoliploides. Além disso, os desenvolvimentos teóricos apresentados neste trabalho abrem caminhos para o adequado estudo da genética de espécies poliploides de maneira geral.

REFERÊNCIAS

AITKEN, K.S.; JACKSON, P.A.; MCINTYRE, C.L. A combination of AFLP and SSR markers provides extensive map coverage and identification of homo(eo)logous linkage groups in a sugarcane cultivar. **Theoretical and Applied Genetics**, New York, v. 110, p. 789–801, 2005.

AITKEN, K.S.; JACKSON, P.A.; MCINTYRE, C.L. Construction of a genetic linkage map for Saccharum officinarum incorporating both simplex and duplex markers to increase genome coverage. **Genome**, Ottawa, v. 50, p. 742–756, 2007.

ALWALA, S.; KIMBENG, C.A.; VEREMIS, J.C.; GRAVOIS, K.A. Linkage mapping and genome analysis in a Saccharum interspecific cross using AFLP, SRAP and TRAP markers. **Euphytica**, Wageningen, v. 164, p. 37–51, 2008.

ANDRU, S.; PAN, Y.B.; THONGTHAWEE, S.; BURNER, D.M.; KIMBENG, C.A. Genetic analysis of the sugarcane (*Saccharum* spp.) cultivar 'LCP 85-384'. I. Linkage mapping using AFLP, SSR, and TRAP markers. **Theoretical and Applied Genetics**, New York, v. 123, p. 77–93, 2011.

APPELS, R.; MORRIS, R.; GILL, B.S.; MAY, C.E. **Chomosome Biology.** Boston: Kluwer Academic, 1998. 424 p.

AUGER, D.L.; GRAY, A.D.; REAM, T.S.; KATO, A.; COE, E.H.; BIRCHLER, J.A. Nonadditive gene expression in diploid and triploid hybrids of maize. **Genetics**, Bethesda, v. 169, p. 389–397, 2005.

BAKER, P.; JACKSON, P.; AITKEN, K. Bayesian estimation of marker dosage in sugarcane and other autopolyploids. **Theoretical and Applied Genetics**, New York, v. 120, p. 1653–72, 2010.

BAUM, E.; PETRIE, T.; G., S.; N., W. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. **The Annals of Mathematical Statistics**, Ann Harbor, v. 41, p. 164–171, 1970.

BIELIG, L.M.; MARIANI, A.; BERDING, N. Cytological studies of 2n male gamete formation in sugarcane. **Euphytica**, Wageningen, v. 133, p. 117–124, 2003.

BIRCHLER, J.A.; AUGER, D.L.; RIDDLE, N.C. In search of the molecular basis of heterosis. **The Plant Cell**, Baltimore, v. 15, p. 2236–2239, 2003.

BOTSTEIN, D.; WHITE, R.L.; SKOLNICK, M.; DAVIS, R.W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. **American Journal of Human Genetics**, Chicago, v. 32, p. 314–331, 1980.

BROMAN, K. W.; SEN, S. **A Guide to QTL Mapping with R/qtl**. New York: Springer, 2009. 400 p.

BROMAN, K.W. **Genetic map construction with R/qtl**. Disponível em: <www.rqtl.org/tutorials/geneticmaps.pdf>. Acesso em: 20 fev. 2010.

BURNHAM, C. R. **Discussions in Cytogenetics**. Mineapolis: Burgess Pub. Co, 1962. 375 p.

CAO, D.; CRAIG, B.A.; DOERGE, R.W. A Model Selection-Based Interval-Mapping Method for Autopolyploids. **Genetics**, Bethesda, v. 169, p. 2371–2382, 2005.

CAO, D.; OSBORN, T.C.; DOERGE, R.W. Correct estimation of preferential chromosome pairing in autotetraploids. **Genome research**, Woodbury, v. 14, p. 459–62, 2004.

COMAI, L. The advantages and disadvantages of being polyploid. **Nature Reviews Genetics**, New York, v. 6, p. 836–846, 2005.

CUADRADO, A.; ACEVEDO, R.; ESPINA, M.D.S.; JOUVE, N.; TORRE, C. Genome remodelling in three modern S. officinarum x S. spontaneum sugarcane cultivars. **Journal of Experimental Botany**, London, v. 55, p. 847–854, 2004.

DA SILVA, J.G.A. **A method for genome mapping of autopolyploids and its application to sugarcane *Saccharum* spp.** 1993. 384 p. Tese (Doutorado) – Cornell University, New York, Ithaca, 1993

DA SILVA, J.A.G.; SORRELLS, M.E. Linkage analysis in polyploids using molecular markers. In: JAUHAR, p. (Ed.) **Methods of Genome Analysis in Plants**. Boca Raton: CRC Press, 1996. cap. 13, p. 211–228.

DEMPSTER, A.P.; LAID, N.M.; RUBIN, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. **Journal of the Royal Statistical Society, Series B**, London, v. 39, p. 1–38, 1977.

D'HONT, A.; GRIVET, L.; FELDMANN, P.; RAO, S.; BERDING, N.; GLASZMANN, J.C. Characterisation of the double genome structure of modern sugarcane cultivars (*Saccharum* spp.) by molecular cytogenetics. **Molecular And General Genetics**, Berlin, v. 250, p. 405–413, 1996.

D'HONT, A.; ISON, D.; ALIX, K.; ROUX, C.; GLASZMANN, J.C. Determination of basic chromosome numbers in the genus *Saccharum* by physical mapping of ribosomal RNA genes. **Genome**, Ottawa, v. 41, p. 221–225, 1998.

D'HONT, A. Unraveling the genome structure of polyploids using FISH and GISH; examples of sugarcane and banana. **Cytogenetic and genome research**, Basel, v. 109, p. 27–33, 2005.

DOERGE, R.W. Constructing Genetic Maps By Rapid Chain Delineation. **Journal of Quantitative Trait Loci**, Washington, v. 2, p. 1–14, 1996.

DOERGE, R.W.; CRAIG, B.A. Model selection for quantitative trait locus analysis in polyploids. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, v. 97, p. 7951-7956, 2000.

EDME, S.J.; GLYNN, N.G.; COMSTOCK, J.C. Genetic segregation of microsatellite markers in *Saccharum* officinarum and S. spontaneum. **Heredity**, London, v. 97, p. 366–375, 2006.

FALK, C.T. A simple scheme for preliminary ordering of multiple loci: application to 45 CF families. In: ELSTON, R.C.; SPENCE, M.A.; HODGE, S.E.; MACCLUER, J.W. (Eds.). **Multipoint Mapping and Linkage based upon Affected Pedigree Members.**, New York: Alan R Liss, 1989. cap. 6, p. 17-22.

FISCHER, R.A. The theory of linkage in polysomic inheritance. **Philosophical Transactions of the Royal Society of London: Series B**, London, v. 23, p. 55–87, 1947.

GALITSKI, T.; SALDANHA, A.J.; STYLES, C.A.; LANDER, E.S.; FINK, G.R. Ploidy Regulation of Gene Expression. **Science**, Washington, v. 285, p. 251–254, 1999.

GALLAIS, A. **Quantitative genetics and breeding methods in autopolyploids plants**. 1.ed. Paris: INRA, 2003. 522 p.

GARCIA, A.A.F.; KIDO, E.A.; MEZA, A.N.; SOUZA, H.M.B.; PINTO, L.R.; PASTINA, M.M.; LEITE, C.S.; SILVA, J.A. G.D.; ULIAN, E.C.; FIGUEIRA, A.V.; SOUZA, A.P.; SILVA, J.A.G.; ULIAN, E.C.; FIGUEIRA, A.V.; SOUZA, A.p. Development of an integrated genetic map of a sugarcane (*Saccharum* spp.) commercial cross, based on a maximum-likelihood approach for estimation of linkage and linkage phases. **Theoretical and Applied Genetics**, New York, v. 112, p. 298–314, 2006.

GAZAFFI, R. **Desenvolvimento de modelo genético-estatístico para mapeamento de QTLs em progênie de irmãos completos, com aplicação em cana-de-açúcar**. 2011. 104 p. Tese (Doutorado em Genética e Melhoramento de Plantas) – Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2011.

GAZAFFI, R.; OLIVEIRA, K.M.; SOUZA, A.P.; GARCIA, A.A.F. Sugarcane: Breeding Methods and Genetic Mapping. In: CORTEZ, L.A.B. (Ed.) **Sugarcane Bioethanol:** R&D for Productivity and Sustainability. São Paulo: Blucher, 2010. p. 333–344

GRATTAPAGLIA, D.; SEDEROFF, R. Genetic Linkage Maps of Eucalyptus grandis and Eucalyptus urophylla Using a Pseudo-Testcross: Mapping Strategy and RAPD Markers. **Genetics**, Bethesda, v. 1137, p. 1121–1137, 1994.

GRIFFITHS, A.J.F.; GELBART.W.M.; LEWONTIN, R.C.; WESSLER, S.R.; SUZUKI, D.T.; MILLER, J.H. **An Introduction to Genetic Analysis**. 8.ed. New York: W.H. Freeman & Company, 2004. 800 p.

GRIVET, L.; ARRUDA, p. Sugarcane genomics: depicting the complex genome of an important tropical crop. **Current Opinion in Plant Biology**, London, v. 5, p. 122–127, 2001.

GUIMARÃES, C.T. G.; SILLS, G.R.; SOBRAL, B.W.S. Comparative mapping of Andropogoneae: *Saccharum* L. (sugarcane) and its relation to sorghum and maize. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, v. 94, p. 14261–14266, 1997.

GUO, M.; DAVIS, D.; BIRCHLER, J.A. Dosage effects on gene expression in a maize ploidy series. **Genetics**, Bethesda, v. 142, p. 1349–1355, 1996.

HACKETT, C.A. A comment on Xie and Xu : “Mapping quantitative trait loci in tetraploid species”. **Genetical Research**, London, v. 78, p. 187–189, 2001.

HACKETT, C.A.; BRADSHAW, J.E.; MCNICOL, J.W. Interval mapping of quantitative trait loci in autotetraploid species. **Genetics**, Bethesda, v. 159, p. 1819–1832, 2001.

HALDANE, J. The combination of linkage values, and the calculation of distance between linked factors. **Journal of Genetics**, London, v. 8, p. 299–309, 1919.

HALDANE, J. B. S. Theoretical genetics of autopolyploids. **Journal of Genetics**, London, v. 22, p. 359–372, 1930.

HELLER-USZYNSKA, K.; USZYNSKI, G.; HUTTNER, E.; EVERS, M.; CARLIG, J.; CAIG, V.; AITKEN, K.; JACKSON, P.; PIPERIDIS, G.; COX, M.; GILMOUR, R.; D'HONT, A.; BUTTERFIELD, M.; GLASZMANN, J.-C.; KILIAN, A. Diversity Arrays Technology effectively reveals DNA polymorphism in a large and complex genome of sugarcane. **Molecular Breeding**, Berlin, v. 28, p. 37–55, 2010.

HIETER, P.; GRIFFITHS, T. Polyploidy–More Is More or Less. **Science**, Washington, v. 285, p. 210–211, 1999.

HOARAU, J.-Y.; OFFMANN, B.; D'HONT, A.; ROQUES, D.; RISTERUCCI A.M.; GLASZMANN, J.-C.; GRIVET, L. Genetic dissection of a modern sugarcane cultivar (*Saccharum* spp.). I. Genome mapping with AFLP markers. **Theoretical and Applied Genetics**, New York, v. 103, p. 84–97, 2001.

JANNOO, N.; GRIVET, L.; DAVID, J.; D'HONT, A.; GLASZMANN, J.-C. Differential chromosome pairing affinities at meiosis in polyploid sugarcane revealed by molecular markers. **Heredity**, London, v. 93, p. 460–467, 2004.

JIANG, C.; ZENG, Z-B. Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. **Genetica**, London, v. 101, p. 47–58, 1997.

KAO, C.; ZENG, Z-B.; TEASDALE, R.D. Multiple Interval Mapping for Quantitative Trait Loci. **Genetics**, Bethesda, v. 125, p. 1203–1216, 1999.

KOSAMBI, D. The estimation of map distance from recombination values. **Annals of Eugenics**, London, v. 12, p. 172–175, 1944.

LABORDA, P.R.; OLIVEIRA, K.M.; GARCIA, A.A.F.; PATERNIANI, M.E.A.G.Z.; SOUZA, A.p. Tropical maize germplasm: what can we say about its genetic diversity in the light of molecular markers? **Theoretical and Applied Genetics**, New York, v. 111, p. 1288–1299, 2005.

LANDER, E.S.; GREEN, p. Construction of multilocus genetic linkage maps in humans. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, v. 84, p. 2363–2367, 1987.

LANDER, E.S.; GREEN, P.; ABRAHAMSON, J.; BARLOW, A.; DALY, M.J.; LINCOLN, S.E.; NEWBERG, L.A. MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. **Genomics**, Orlando, v. 1, p. 174–181, 1987.

LEACH, L.J.; WANG, L.; KEARSEY, M.J.; LUO, Z. Multilocus tetrasomic linkage analysis using hidden Markov chain model. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, v. 107, p. 4270–4274, 2010.

LEWIN, H.A.; LARKIN, D.M.; PONTIUS, J.; O'BRIEN, S.J.; BRIEN, S.J.O. Every genome sequence needs a good map. **Genome Research**, Woodbury, v. 19, p. 1925–1928, 2009.

LIU, B. **Statistical Genomics: Linkage, Mapping, and QTL Analysis**. Boca Raton: CRC Press, 1998.

LUO, Z.W.; TAO, S.H.; ZENG, Z-B. Inferring Linkage Disequilibrium Between a Polymorphic Marker Locus and a Trait Locus in Natural Populations. **Genetics**, v. 156, p. 457-467, 2000.

LUO, Z.W.; HACKETT, C.A.; BRADSHAW, J.E.; MCNICOL, J.W.; MILBOURNE, D. Predicting parental genotypes and gene segregation for tetrasomic inheritance. **Theoretical and Applied Genetics**, New York, v. 100, p. 1067–1073, 2000.

LUO, Z.W.; HACKETT, C.A.; BRADSHAW, J.E.; MCNICOL, J.W.; MILBOURNE, D. Construction of a genetic linkage map in tetraploid species using molecular markers. **Genetics**, Bethesda, v. 157, p. 1369–1385, 2001.

LUO, Z.W.; ZHANG, R.M.; KEARSEY, M.J. Theoretical basis for genetic linkage analysis in autotetraploid species. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, v. 101, p. 7040–7045, 2004.

LUO, Z.W.; ZHANG, Z.; LEACH, L.; ZHANG, R.M.; BRADSHAW, J.E.; KEARSEY, M.J. Constructing genetic linkage maps under a tetrasomic model. **Genetics**, Bethesda, v. 172, p. 2635–2645, 2006.

MA, C-X.; CASELLA, G.; SHEN, Z-J.; OSBORN, T.C.; WU, R. A Unified Framework for Mapping Quantitative Trait Loci in Bivalent Tetraploids Using Single-dose Restriction Fragments: A Case Study from Alfalfa. **Genome research**, Woodbury, v. 12, p. 1974–1981, 2002.

MARGARIDO, G.R.A.; SOUZA, A.P.; GARCIA, A.A.F. OneMap: software for genetic mapping in outcrossing species. **Hereditas**, Lund, v. 144, p. 78–79, 2007.

MARGARIDO, G.R.A.; MOLLINARI, M.; GARCIA, A.A.F. **OneMap Tutorial**. Disponível em: <http://cran.r-project.org/web/packages/onemap/vignettes/>. Acesso em: 20 fev. 2010.

MATHER, K. Segregation and linkage in autotetraploids. **Journal Genetics**, London, v. 30, p. 287–314, 1936.

MESTER, D.; RONIN, Y.; MINKOV, D.; NEVO, E.; KOROL, A. Constructing large-scale genetic maps using an evolutionary strategy algorithm. **Genetics**, Bethesda, v. 165, p. 2269–2282, 2003.

MEUWISSEN, T.H.E.; HAYES, B.J.; GODDARD, M.E. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. **Genetics**, Bethesda, v. 157, p. 1819–1829, 2001.

MOLLINARI, M. **Comparação de algoritmos usados na construção de mapas genéticos**. 2009. 73 p. Dissertação(Mestrado em Genética e Melhoramento de Plantas) – Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2009.

MOLLINARI, M.; MARGARIDO, G.R.A.; VENCOVSKY, R.; GARCIA, A.A.F. Evaluation of algorithms used to order markers on genetic maps. **Heredity**, London, v. 103, p. 494–502, 2009.

MORGAN, T.H. **The Theory of Genes**. New Haven: Yale University Press, 1928.

MORTON, N.E. Sequential tests for the detection of linkage. **American Journal of Human Genetics**, Baltimore, v. 7, p. 277–318, 1955.

OETH, P.; BEAULIEU, M.; PARK, C.; KOSMAN, D.; MISTRO, G. del; van Den Boom, D.; JURINKE, C. **iPLEX™ Assay:** Increased Plexing Efficiency and Flexibility for MassARRAY® System Through Single Base Primer Extension with Mass-Modified Terminators. San Diego: Sequenom, 2007. 12 p.

OLIVEIRA, K.M.; LABORDA, P.R.; GARCIA, A.A.F.; PATERNIANI, M.E.A.G.Z.; SOUZA, A.p. Evaluating genetic relationships between tropical maize inbred lines by means of AFLP profiling. **Hereditas**, Lund, v. 140, p. 24–33, 2004.

OLIVEIRA, K.M.; PINTO, L.R.; MARCONI, T.G.; MARGARIDO, G.R.A.; PASTINA, M.M.; TEIXEIRA, L.H.M.; FIGUEIRA, A.V.; ULIAN, E.C.; GARCIA, A.A.F.; SOUZA, A.p. Functional integrated genetic linkage map based on EST-markers for a sugarcane (*Saccharum* spp.) commercial cross. **Molecular Breeding**, Berlin, v. 20, p. 189–208, 2007.

OLIVEIRA, E.J.; VIEIRA, M.L.C.; GARCIA, A.A.F.; MUNHOZ, C.F.; MARGARIDO, G.R.A.; CONSOLI, L.; MATTA, F.P.; MORAES, M.C.; ZUCCHI, M.I.; FUNGARO, M.H.p. An Integrated Molecular Map of Yellow Passion Fruit Based on Simultaneous Maximum-Likehood Estimation of Linkage and Linkage Phases. **Journal of the American Society for Horticultural Sciences**, Wageningen, v. 133, p. 35–41, 2008.

OSBORN, T.C.; PIRES, J.C.; BIRCHLER, J.A.; AUGER, D.L.; CHEN, Z.J.; LEE, H-S.; COMAI, L.; MADLUNG, A.; DOERGE, R.W.; COLOT, V.; MARTIENSSEN, R.A. Understanding mechanisms of novel gene expression in polyploids. **Trends in Genetics**, Kidlington, v. 19, n. 3, p. 141–147, 2003.

PASTINA, M.M.; PINTO, L.R.; OLIVEIRA, K.M.; SOUZA, A.P.; GARCIA, A.A.F. Molecular Mapping of Complex Traits. In: HENRY, R.J.; KOLE, C. (Eds.) **Genetics, Genomics and Breeding of Sugarcane**. 1.ed. New Hampshire: Science Publishers, 2010. p. 117-148.

PASTINA, M.M.; MALOSETTI, M.; GAZAFFI, R.; MOLLINARI, M.; MARGARIDO, G. R.A.; OLIVEIRA, K.M.; PINTO, L.R.; SOUZA, A.P.; EEUWIJK, F.A. van; GARCIA, A.A.F. A mixed model qtl analysis for sugarcane multiple-harvest-location trial data. **Theoretical and Applied Genetics**, New York, v. 124, p. 835–849, 2012.

PIPERIDIS, G.; PIPERIDIS, N.; D'HONT, A. Molecular cytogenetic investigation of chromosome composition and transmission in sugarcane. **Molecular Genetics and Genomics**, Berlin, v. 284, p. 65–73, 2010.

PIPERIDIS, N.; PIPERIDIS, G.; D'HONT, A. Molecular Cytogenetics In: HENRY, R.J.; KOLE, C. (Eds.) **Genetics, Genomics and Breeding of Sugarcane**. 1.ed. New Hampshire: Science Publishers, 2010. p. 9-18.

R DEVELOPMENT CORE TEAM (2011). **R:** A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria, 2011. Disponível em: <http://www.r-project.org/>. Acesso em: 10 set. 2011.

RABINER, L. R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. **Proceedings of the Institute of Electrical and Electronics Engineers**, London, v. 77, p. 257-286, 1989.

RIPOL, M.I.; CHURCHILL, G.A.; SILVA,J.A.G.D.; SORRELLS, M.; SILVA, J. A.G. da; SORRELLS, M. Statistical aspects of genetic mapping in autopolyploids. **Gene**, Amsterdam, v. 235, p. 31–41, 1999.

SEQUENOM. **Typer 4.0 Manual**. San Diego: Sequenom, 2007. 179 p.

SERANG, O.; MOLLINARI, M.; GARCIA, A.A.F. Efficient Exact Maximum a Posteriori Computation for Bayesian SNP Genotyping in Polyploids. **PloS One**, v. 7, p. 1–12, 2012.

SINGH, R. J. **Plant Cytogenetics**. 2.ed. Boca Raton: CRC Press, 2003. 463 p.

SOLTIS, D.E.; SOLTIS, P.S. Polyploidy: recurrent formation and genome evolution. **Trends Ecology Evolution**, Cambridge, v. 14, p. 348–352, 1999.

SOUZA, A.p. Biologia Molecular Aplicada ao Melhoramento. In: NASS, L.L.; VALOIS, A.C.C.; MELO, I.S.; VALADARES-INGLIS, M.C. (Eds.) **Recursos Genéticos e Melhoramento – Plantas**. Rondonópolis: Fundação MT, 2001. p. 939–965

SVED, J.A. The Relationship between diploid and tetraploid recombination frequencies. **Heredity**, London, v. 14, p. 348–352, 1962.

SYBENGA, J. **General cytogenetics**. New York: American Elsevier, 1972. 359 p.

SYVÄNEN, A.C. Accessing genetic variation: genotyping single nucleotide polymorphisms. **Nature Reviews Genetics**, London, v. 2, p. 930–42, 2001.

TAN, Y.; FU, Y. A novel method for estimating linkage maps. **Genetics**, Bethsda, v. 173, p. 2383–2390, 2006.

TAUTZ, D. Hypervariability of simple sequences as a general source for polymorphic DNA markers. **Nucleic Acids Research**, London, v. 17, p. 6463–6471, 1989.

VOORRIPS, R.E.; GORT, G.; VOSMAN, B. Genotype calling in tetraploid species from bi-allelic marker data using mixture models. **BMC Bioinformatics**, London, v. 12, p. 172, 2011.

WANG, J.; TIAN, L.; LEE, H.-S.; WEI, N.E.; JIANG, H.; WATSON, B.; MADLUNG, A.; OSBORN, T.C.; DOERGE, R.W.; COMAI, L.; CHEN, Z.J. Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. **Genetics**, Bethesda, v. 172, p. 507–517, 2006.

WANG, J.; ROE, B.; MACMIL, S.; YU, Q.; MURRAY, J.E.; TANG, H.; CHEN, C.; NAJAR, F.; WILEY, G.; BOWERS, J.; VAN SLUYS, M.; ROKHSAR, D.S.; HUDSON, M.E.; MOOSE, S.P.; PATERSON, A.H.; MING, R. Microcollinearity between autopolyploid sugarcane and diploid sorghum genomes. **BMC Genomics**, London, v. 11, p. 261–288, 2010.

WEEKS, D.E.; LANGE, K. Preliminary Ranking Procedures for Multilocus Ordering. **Genomics**, Orlando, v. 1, p. 236–242, 1987.

WILLIAMS, J.G.K.; KUBELIK, A.R.; LIVAK, K.J.; RAFALSKI, J.A.; SCOTT,V. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. **Nucleic Acids Research**, London, v. 18, p. 6531-6535, 1990.

WILSON, S. A major simplification in the preliminary ordering of linked loci. **Genetic Epidemiology**, New York, v. 5, p. 75–80, 1988.

WOLFE, K. H. Yesterday's polyploids and the mystery of diploidization. **Nature Review Genetics**, London, v. 2, p. 333–341, 2001.

WU, K.K.; BURNQUIST, W.; SORRELLS, M.E.; TEW, T.L.; MOORE, P.H.; TANKSLEY, S. D. The detection and estimation of linkage in polyploids using single-dose restriction fragments. **Theoretical and Applied Genetics**, New York, v. 83, p. 294–300, 1992.

WU, R.; GALLO-MEAGHER, M.; LITTELL, R.C.; ZENG, Z. A General Polyploid Model for Analyzing Gene Segregation in Outcrossing Tetraploid Species. **Genetics**, Bethesda, v. 159, p. 869–882, 2001.

WU, R.; MA, C.; PAINTER, I.; ZENG, Z. Simultaneous Maximum Likelihood Estimation of Linkage and Linkage Phases in Outcrossing Species. **Theoretical Population Biology**, New Yourk, v. 61, p. 349–363, 2002.

WU, R.; MA, C.; WU, S.S.; ZENG, Z. Linkage mapping of sex-specific differences. **Genetic Research**, Oxford, v. 79, p. 85–96, 2002.

VOS, P.; HOGERS, R.; BLEEKER, M.; REIJANS, M.; VANDELEE, T.; HORNES, M.; FRIJTERS, A.; POT, J.; PELEMAN, J.; KUIPER, M.; ZABEAU, M. AFLP: a new technique for DNA fingerprinting. **Nucleic Acids Research**, Oxford, v. 23, p. 4407-4414, 1995.

ZENG, Z.; KAO, C.; BASTEN, C.J. Estimating the genetic architecture of quantitative traits. **Genetical Research**, Oxford, v. 74, p. 279–289, 1999.

Apêndice

APÊNDICE 1 - Simplificações das probabilidades de transição

$$\mathrm{P}(\mathbf{p}_{k+1}^m|\mathbf{p}_k^m) = \frac{\mathrm{P}(\mathbf{p}_k^m,\ \mathbf{p}_{k+1}^m)}{\sum_{l=0}^{\frac{m}{2}} \mathrm{P}(\mathbf{p}_k^m,\ \mathbf{p}_{k+1}^m)\binom{\frac{m}{2}}{l}^2}$$

$$\mathrm{P}(\mathbf{p}_k^m,\ \mathbf{p}_{k+1}^m) = \frac{\frac{(1-r_k)^{\frac{m}{2}-l}(r_k)^l}{2^{\frac{m}{2}}} l!(\frac{m}{2}-l)!}{\frac{1}{\frac{m}{2}!}\prod_{i=2,4,\cdots,m}\binom{i}{2}}$$

$$= (1-r_k)^{\frac{m}{2}-l}(r_k)^l \frac{l!(\frac{m}{2}-l)!\frac{m}{2}!}{2^{\frac{m}{2}}\prod_{i=2,4,\cdots,m}\binom{i}{2}} \quad (11)$$

$$\sum_{l=0}^{\frac{m}{2}} \mathrm{P}(\mathbf{p}_k^m,\ \mathbf{p}_{k+1}^m)\binom{\frac{m}{2}}{l}^2 = \sum_{l=0}^{\frac{m}{2}} \binom{\frac{m}{2}}{l}^2 \frac{\frac{(1-r_k)^{\frac{m}{2}-l}(r_k)^l}{2^{\frac{m}{2}}} l!(\frac{m}{2}-l)!}{\frac{1}{\frac{m}{2}!}\prod_{i=2,4,\cdots,m}\binom{i}{2}}$$

$$= \sum_{l=0}^{\frac{m}{2}} \binom{\frac{m}{2}}{l}(1-r_k)^{\frac{m}{2}-l}(r_k)^l \frac{l!(\frac{m}{2}-l)!\frac{m}{2}!}{2^{\frac{m}{2}}\prod_{i=2,4,\cdots,m}\binom{i}{2}} \frac{\frac{m}{2}!}{l!(\frac{m}{2}-l)!}$$

$$= \frac{\frac{m}{2}!^2}{2^{\frac{m}{2}}\prod_{i=2,4,\cdots,m}\binom{i}{2}} \sum_{l=0}^{\frac{m}{2}} \binom{\frac{m}{2}}{l}(1-r_k)^{\frac{m}{2}-l}(r_k)^l$$

$$= \frac{\frac{m}{2}!^2}{2^{\frac{m}{2}}\prod_{i=2,4,\cdots,m}\binom{i}{2}} \{(1-r_k)r_k\}^{\frac{m}{2}}$$

$$= \frac{\frac{m}{2}!^2}{2^{\frac{m}{2}}\prod_{i=2,4,\cdots,m}\binom{i}{2}} \quad (12)$$

Dividindo-se (11) por (12)

$$\begin{aligned} \mathrm{P}(\mathbf{p}_{k+1}^m|\mathbf{p}_k^m) &= (1-r_k)^{\frac{m}{2}-l}(r_k)^l \frac{l!(\frac{m}{2}-l)!}{\frac{m}{2}!} \\ &= \frac{(1-r_k)^{\frac{m}{2}-l}(r_k)^l}{\binom{\frac{m}{2}}{l}} \end{aligned} \tag{13}$$